分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2210330179

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于深度信息融合的跨模态行人重识别方法研究

论文作者　王路遥
校内导师　王凤随 教授
校外导师　徐文霞 工程师

专业学位类别　电子信息
研究方向（领域）　智能信息处理及应用

论文提交日期：2024 年 5 月 31 日

分类号：TP391　　单位代码：10363

密　级：公 开　　学　号:2210330179

题　目 基于深度信息融合的跨模态行人重识别方法研究

题　目 Research on Cross-Model Person Re-Identification Method Based on Deep Information Fusion

学　生　姓　名：王路遥

校　内　导　师：王凤随（教授）

校　外　导　师：徐文霞（工程师）

专　　业：电子信息

研　究　方　向：智能信息处理及应用

论文答辩日期：2024 年 5 月 23 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：王路遥

日期： 2024 年 5 月 26 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：王路遥　　指导教师签名：王凤随

日期：　2024 年 5 月 26 日　　日期：　2024 年 5 月 26 日

基于深度信息融合的跨模态行人重识别方法研究

摘 要

随着监控系统在城市地区的快速发展，行人重识别作为智能监控、家居安防、公安刑侦等领域的核心技术引起了广泛关注。目前，基于深度学习的行人重识别技术在可见光环境下取得了显著的进展。然而，在夜间、弱光等条件下，可见光摄像机无法捕捉到有效的行人外观特征，并且智能监控全天候运作，需要对白天捕获的可见光图像和夜间捕获的红外图像同时进行处理，增大了查人寻踪的难度。因此，提出针对可见光-红外的跨模态行人重识别研究。该任务的关键是可见光图像和红外图像之间的跨模态行人检索匹配问题，一方面考虑如何处理行人由遮挡、各种姿态等引起的模态内差异；另一方面考虑如何解决图像间的跨模态差异。论文主要围绕不同模态图像间的深度特征信息展开，为此，提出基于深度信息融合的跨模态行人重识别改进方法，其主要研究内容如下：

（1）跨模态行人重识别大部分工作主要集中于调整不同模态的特征分布来弱化模态差异，往往忽略了行人的细节信息和网络中噪声的干扰。针对这一问题，提出一种深度信息融合变分细化蒸馏的跨模态行人重识别改进方法。该方法首先利用多分支聚合不同粒度的深度全局特征，督促深层网络学习两种模态的全局信息和细节信息，丰富行人的特征描述符；然后，结合变分细化蒸馏策略，对特

征信息进行再压缩，保留与任务相关的深层信息，同时丢弃无用的干扰物；最后将网络捕获的不同特征用多种损失函数联合监督，以提高网络对行人表征的敏感度。在跨模态行人重识别数据集 SYSU-MM01 的全搜索模式下所提方法 R-1 达到 56.36%、mAP 达到 54.96%、mINP 达到 42.01%。实验结果证明了所提方法的有效性。

（2）由于跨模态的图像数据分布差距大，使其难以在不同模态下进行行人的身份匹配。故针对现有网络模型中训练数据不充足以及跨模态之间较大模态差异的问题，提出一种多样化深度信息融合嵌入的跨模态行人重识别改进方法。该方法通过多分支生成多样化的深度特征信息嵌入，并与原始输入的特征图进行融合拼接，拓展训练数据，获取丰富的语义信息。其次，为减少模态内和模态间差异，提出一种异质中心共享三元组损失对行人的类内类间距离进行约束，督促网络学习与模态无关且能够有效扩大类间差异缩小类内紧凑型的特征。最后与交叉熵损失共同作用于网络，提升模型的分类能力。在两个公共数据集下的实验结果证明了所提方法的有效性。其中，在 SYSU-MM01 数据集的全搜索模式下，所提方法 R-1 达到 69.98%、mAP 达到 66.71%、mINP 达到 52.83%。

（3）跨模态行人图像间存在遮挡、错位及颜色变化导致的巨大差异，降低了网络学习行人特征表示和处理跨模态差异的能力。针对这些问题，提出一种融合多尺度特征与混淆学习的跨模态行人重识别改进方法。为实现高效的特征提取，缩小模态内差异，将网络

设计为多尺度特征互补的形式，分别学习行人的局部细化特征与全局粗糙特征，从细粒度和粗粒度两方面来增强网络的特征表达能力。利用混淆学习策略，将真伪标签与特征信息融合以模糊网络的模态识别反馈，挖掘稳定且有效的模态无关属性应对模态差异，来提高特征对模态变化的鲁棒性。在大规模数据集 SYSU-MM01 的全搜索模式下所提算法 R-1 和平均精度(mAP)的结果分别为 76.69%和 72.45%，在 RegDB 数据集的可见光到红外模式下，所提方法 R-1 和平均精度(mAP)的结果分别为 94.62%和 94.60%，优于现有的主流方法，验证了所提方法的有效性。

关键词：行人重识别，信息融合，跨模态，多分支，多损失联合

RESEARCH ON CROSS-MODEL PERSON RE-IDENTIFICATION METHOD BASED ON DEEP INFORMATION FUSION

ABSTRACT

With the rapid development of monitoring systems in urban areas, person re-identification has attracted wide attention as the core technology in the fields of intelligent monitoring, home security, public security and criminal investigation. At present, person re-identification technology based on deep learning has made significant progress in visible light environments. However, under conditions such as nighttime and low light, visible light cameras are unable to capture effective pedestrian appearance features, and intelligent monitoring operates all-weather, requiring simultaneous processing of visible images captured during the day and infrared images captured at night, increasing the difficulty of human tracking. Therefore, infrared-visible cross-modal person re-identification research is proposed. The key to this task is the cross-modal person retrieval matching problem between visible images and infrared images. On the one hand, it considers how to deal with the intra-modal differences caused by occlusion and various poses. On the other hand, consider how to solve the cross-modal differences between images. The paper mainly focuses on the deep feature information between different modal images. Therefore, an improved method for cross-modal person re-identification based on deep information fusion is proposed. The main research contents are as follows:

(1) Most of the cross-modal person re-identification work focuses on adjusting the feature distribution of different modalities to weaken modal differences, often ignoring the interference of pedestrian details and noise in the network. A cross-modal person re-identification improvement method based on deep information fusion variational refinement distillation is proposed to address this issue. This method first utilizes multi branch aggregation of deep global features with different granularities to urge the deep network to learn the global and detailed information of two modalities, enriching the feature descriptors of pedestrians; Then, combined with the variational refinement distillation strategy, the feature information is recompressed to retain the deep information related to the task while discarding useless interference. Finally, the different features captured by the network are jointly supervised by multiple loss functions to improve the sensitivity of the network to pedestrian representation. In the all-search mode of the cross-modal person re-identification dataset SYSU-MM01, the proposed method R-1 reaches 56.36%, mAP reaches 54.96%, and mINP reaches 42.01%. The experimental results demonstrate the effectiveness of the proposed method.

(2) Due to the large distribution difference of cross modal image data, it is difficult to perform person identity matching in different modalities. Therefore, in response to the problem of insufficient training data and significant modal differences between different modalities in existing network models, an improved method for cross-modal person re-identification based on diversified feature information fusion embedding is proposed. This method generates diverse deep feature information embeddings through multiple branches and fuses them with the original input feature map to expand the training data and obtain rich semantic information. Secondly, to reduce intra-modal and inter-modal differences,

a heterogeneous center-shared triplet loss is proposed to constrain the intra-class and inter-class distances of pedestrians, urging the network to learn features that are independent of modalities and can effectively expand inter-class differences and reduce the intra-class compact features. Finally, the network is combined with the cross entropy loss to improve the classification ability of the model. The experimental results on two public datasets demonstrate the effectiveness of the proposed method. Among them, in the all-search mode of the SYSU-MM01 dataset, the proposed method R-1 achieved 69.98%, mAP reached 66.71%, and mINP reached 52.83%.

(3) There are huge differences caused by occlusion, misalignment and color changes between cross-modal person images, which reduces the ability of the network to learn pedestrian feature representation and deal with cross-modal differences. A cross-modal person re-identification improvement method that integrates multi-scale features and confusion learning is proposed to address these issues. In order to achieve efficient feature extraction and reduce intra-modal differences, the network is designed as a complementary form of multi-scale features. The local refinement features and global rough features of pedestrians are learned respectively, and the feature expression ability of the network is enhanced from fine-grained and coarse-grained aspects. Using a confusion learning strategy to fuse true and false labels with feature information for fuzzy network modal recognition feedback, and stable and effective mode-independent attributes are mined to cope with mode differences, thereby improving the robustness of features to mode changes. The results of the proposed algorithm R-1 and average precision (mAP) in the all-search mode of the large-scale dataset SYSU-MM01 are 76.69% and 72.45%, respectively. In the visible to infrared mode of the RegDB dataset, the

proposed method R-1 and average precision (mAP) are 94.62% and 94.60%, respectively, which are superior to existing methods and verify the effectiveness of the proposed method.

KEY WORDS: person re-identification, information fusion, cross-modal, multi-branch, joint multi-loss

目录

第 1 章 绪论

1.1 研究背景与意义

随着近几年科学技术水平的提升与社会生活的信息化发展，人们的精神文化生活得到了保障，使得人们对公共安全技术的需求持续增长。智能监控系统作为维护城市公共安全的关键，已广泛应用于社区、家居安防、智能交通管控以及数字化城市管理等领域。为防止违法犯罪和意外事件的发生，部署大量摄像头于居民小区、学校、公园、商场等公共场所来预防犯罪及查人寻踪。视频监控成为城市管理和社会治安的重要手段。传统的侦察方式是利用人工翻阅大量视频片段来实现跨摄像头追踪特定的行人，这种方法往往需要耗费大量的时间和精力。因此，能够实现跨摄像头对行人进行追踪定位的智能监控系统显得尤为重要。智能监控系统利用计算机视觉对摄相机采集的视频图像进行智能分析，能够自动的分析和识别关键目标信息，如车辆、行人等，有效提高了对特定目标的检索效率并且降低了漏检、误检率，节约了人防成本。同时，可以快速识别可疑人员和危险物品，防止意外事件发生，提高了社会安全水平。随着深度学习的快速发展，可以跨摄像头跨场景检索相同行人的行人重识别技术应运而生。

行人重识别（Person Re-Identification, Re-ID）作为智能化监控的关键旨在不同时间、地点跨越不同的摄像头来检索目标行人图像，该技术可以帮助监控系统快速准确地识别和追踪目标行人[1-3]。查询行人数据源可以由不同摄像机捕获的视频[4-5]、图像[6-7]或文本描述[8-9]等表示，其中行人图像为本研究的查询数据。由于不同摄像机拍摄行人的视角不同，在现实环境中 Re-ID 研究存在许多困难，如图像背景差异、障碍物遮挡、行人的姿态改变以及摄像机参数不同等多种因素，如图 1-1 所示，使同一个行人的外观差异巨大，从而影响 Re-ID 的检索准确度。

图 1-1 不同摄像机拍摄视角下的同一行人

传统的行人重识别仅针对 RGB 图像，进行单模态行人图像匹配，而在实际的智能监控中，摄像头同时也需要在光线不充足或黑夜场景进行实时监控。由于夜晚的能见度较低，为了保证实际的监控系统能够全天候作用，摄像机需要由可见光（Red Green Blue, RGB）模式转换到红外（Infrared Radiation, IR）模式来捕获行人的外观特征。因此，Re-ID 需要在 RGB 和 IR 图像间进行检索，跨模态行人重识别问题被提出。跨模态 Re-ID 用于匹配同一行人的三通道 RGB 图像和单通道 IR 图像，使其不仅面临传统单模态 Re-ID 的挑战，还面临异质数据源的模态差异难题，如图 1-2 所示。

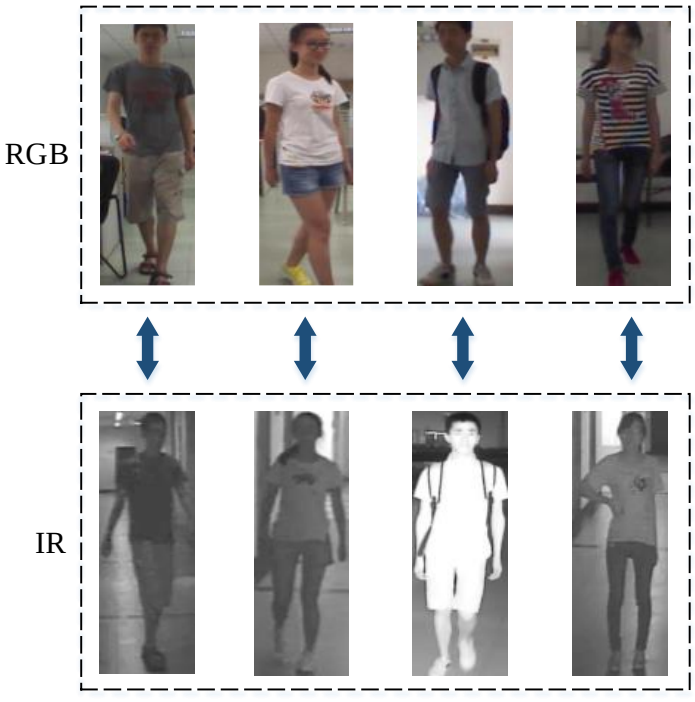


图 1-2 跨模态行人重识别异构数据差异

基于智能监控系统的行人重识别技术流程如图 1-3 所示，主要包括以下几个步骤：（1）原始数据采集：采集不同场景下的行人图像。（2）行人边界框生成：

利用行人检测或跟踪算法对图像中的行人进行定位分割，得到行人区域的图片。（3）训练数据标注：标注跨摄像机标签。（4）模型训练：利用标注的行人图像训练出健壮的 Re-ID 模型。（5）行人检索：给定一个待查询的行人图像和图库集，利用训练后的 Re-ID 模型捕获的行人表征对查询图像和图库图像进行相似度排序。

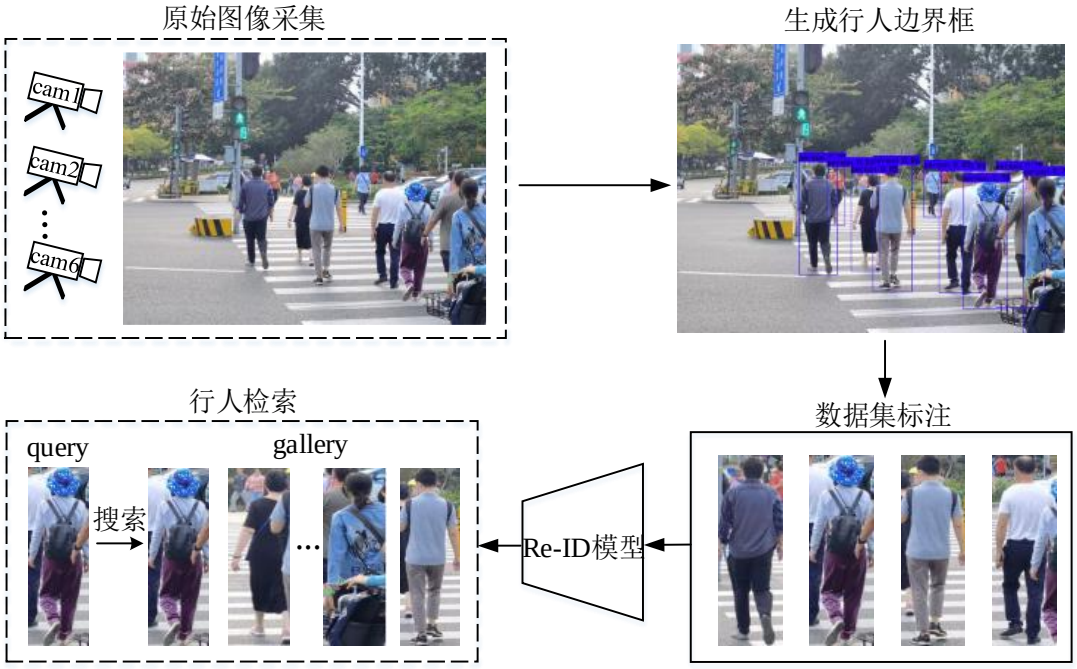


图 1-3 行人重识别系统流程

1.2 国内外研究现状

1.2.1 单模态行人重识别研究现状

行人重识别任务被视为一个多分类问题，将具有相同身份行人图像归属于同一类别，通常用于处理单一行人模态的情况，进行 RGB-RGB 图像的匹配。2012 年，Krizhevsky 等人[10]在 ILSVRC 比赛上以巨大的优势赢得冠军，此后基于卷积神经网络的深度模型在图像处理领域备受欢迎。随着深度学习的发展，基于卷积神经网络的 Re-ID 技术在计算机视觉领域逐渐成为热点。与深度学习结合的 Re-ID 研究可分为基于特征学习的方法[11-13]和基于度量学习的方法[14-20]。

基于特征学习的 Re-ID 重点关注如何更好的学习有鉴别性的高级特征或颜色、纹理等低级特征的语义表示。主要分为基于全局特征、局部特征或辅助特

征学习三类。基于全局特征学习，即学习行人图像中的全局特征进行相似度匹配[21]。如 Wang 等人[22]提出将单图像和交叉图像表示联合学习，优化建模探针图像和图库图像之间的关系，利用三元组损失监督训练，提高图像匹配效率。Si 等人[23]提出一种端到端的双注意力匹配网络，该网络主要由特征提取序列和特征匹配序列的两个模块组成。其中，特征序列提取模块进行特征细化，特征匹配模块致力于实现特征对齐。上述从全局特征角度学习的方法容易受图像中噪声，如背景杂波等区域的干扰，导致特征信息无法正确匹配。为解决这一问题，提出提取图像的局部特征进行学习匹配。基于局部特征学习[24]，即提取图像中行人的部分或多尺度的局部特征用于图像检索匹配，进一步增强行人特征的判别能力。Sun 等人[25]提出一种局部卷积基线网络（Part-based Conv Baseline, PCB），将行人特征图水平分割为 6 个条纹区域以提取局部特征，考虑到水平划分容易导致局部区域包含异常值，提出一种部分精炼池化（Refined Part Pooling, RPP）策略，对异常值进行重新分配来纠正局部不一致性。Wang 等人[26]提出一种整合全局和不同尺度局部特征信息的学习方法，设计了多粒度网络（Multiple Granularity Network, MGN），该网络采用将图像细分为多个条带的策略，并在各个分支中调整条带的数量，从而实现表征的多粒度特性，最终结合不同粒度的全局信息和局部信息以实现更强大的行人表示。Kalayeh 等人[27]提出 SPReID 方法，使用人体分割支路辅助提取行人的局部特征，利用局部视觉获取像素级判别信息进行重识别，以增强模型对行人不同角度、姿态变化的鲁棒性。图像中基于全局特征和局部特征学习虽然能够一定程度上提升的模型识别精度，但仍不能弥补图像受局部遮挡导致信息丢失的问题。基于辅助特征，研究中常借助语义属性[28-29]、生成对抗网络所生成的虚拟图像[30]等辅助信息，在全局和局部特征学习基础上建立图像特征间相关性，以学习更多相关特征信息，进一步增强表征学习能力。Matsukawa 等人[31]提出基于多个行人属性标签的分类损失和组合属性分类损失改进网络，迫使网络学习更多具有区分性的行人特定信息，进一步模型的提高识别能力。Su 等人[32]通过深入学习人类属性特征提出一种半监督的深度属性学习方法，首先利用有标记属性的数据集进行初始训练，接着利用易获得的行人标签细化初始模型，最后将有标记属性的数据集与初始标记的目标数据集组合对网络进行更新训练，增强了网络的属性预测学习和泛化能

力。Zheng 等人[33]采用深度卷积生成对抗网络进行样本生成，将原始的训练数据和生成的样本共同作为网络的输入进行训练，以改进网络的特征表示学习。

与上述特征学习的方法相比，基于度量学习的方法能够使模型更好地学习数据之间的潜在相关性，并通过区分样本之间的特征距离比较来学习具有判别性的特征，进一步提高模型的分类效率。Ding 等人[34]提出了一个可扩展的深度特征学习模型，利用三元组损失拉近嵌入空间中相同身份的行人特征，扩大不同身份的行人特征距离来优化网络学习过程。Hermans[35]等人提出一种难样本挖掘的三元组损失，通过挖掘难样本训练网络，有效提升了模型的泛化能力。随后，Chen 等人[36]设计了一种四元组损失，基于三元组引入了一个新的约束构成四元组，进一步实现了缩小类内变化，增加类间差异的目的。Zhao 等人[37]在三元组损失的基础上进行改进，设计了一种新的硬挖掘中心三元组损失。该损失利用样本类别的中心形成硬三元组来优化行人的类内和类间距离，并且减少了计算和挖掘硬样本的成本。Lou 等人[38]利用身份损失对行人图像进行分类，同时采用标签平滑的方式防止 Re-ID 模型对行人的标签信息过度拟合，来提高网络的泛化能力。

1.2.2 跨模态行人重识别研究现状

单模态的行人重识别进行可见光图像间的匹配，不适应于夜晚能见度较低的情况。为了保证实际的监控系统能够全天候作用，摄像机在夜间自动转换红外模式来记录行人的特征信息。随着 RGB-IR 相机的大量使用，跨模态行人重识别研究受到视觉领域的广泛关注，该研究侧重于处理 RGB 和 IR 图像间的跨模态差异，关注如何获取具有鉴别性的跨模态共享信息，来改善网络模型的性能。目前，跨模态行人重识别方法主要分为：基于全局和局部特征的学习方法、基于度量学习的方法和基于模态转换网络的方法。

在全局特征学习方面，2017 年，Wu 等人[39]提出了深度零填充（Zero-padding）的方法来训练单流网络，通过填充跨模态输入到特定域的节点，以自适应地学习模态共享特征，并且在现实场景下搭建了一个适用于跨模态 Re-ID 研究的大规模数据集 SYSU-MM01。Ye 等人[40]提出双流卷积神经网络（Two-stream CNN Network, TONE），包括特征学习和度量学习两个阶段，通过学习模

态特异性信息与模态共享矩阵来分别处理交叉模态差异和模态内变化，同时整合对比损失函数，缩小模态间的差距，并促进表示学习的模态一致性。随后，Ye 等人[41]又针对跨模态行人重识别研究提出了端到端的双路径学习框架，该框架由一个用于捕获特征的双路径和用来判别特征的双向双约束排序（Bi-directional Dual-constrained Top-Ranking, BDTR）损失构成，以保证网络的识别率。Feng 等人[42]设计了一个端到端的框架，网络采用模态独立的分支分别学习不同模态的图像特征，同时为缩小模态间差异，采用身份损失学习图像中不随模态改变的特征信息，以提高模型的判别性。上述方法从全局特征学习角度提升模型的识别准确度，由于全局特征学习的特征范围较大，易造成信息不对齐和丢失图像中重要的细节信息等问题，故提出对图像某一小部分区域进行提取的局部特征学习。在局部特征学习方面，Zhu 等人[43]构建了双流的局部特征网络（Two-Stream Local Feature Network, TSLFN），RGB 图像和 IR 图像分别输入两条参数不共享的骨干网络，以学习特定模态的特征信息，并且将输出的特征均匀划分为多条条纹进行局部特征学习，接着将来自不同模式的特征投射到同一子空间用共享权值的方式来学习模态共享特征。Ye 等人[44]设计了一个用于区分局部聚合特征学习的模态内加权局部注意力，在模态内挖掘上下文部分关系来学习有区别的部分聚合特征，自适应的分配不同局部部位的权重进行行人重识别。Hao 等人[45]设计一种双对齐特征嵌入方法，利用局部特征代替全局特征提取细粒度的模态不变信息，来提高特征的判别性。Wei 等人[46]提出一种基于灵活身体划分模型的对抗性学习（Flexible Body Partition model-based Adversarial Learning, FBP-AL）方法，引入自适应加权学习来自动分割图像中行人的身体部位，根据重要性实现对各个部分的学习，使网络能够更多学习具有区别的细粒度表示的部分。

在深度度量学习方面，通过度量学习函数将不同的行人图像映射到相同的特征空间。利用模型提取特征，计算特征之间的距离，在特征空间中拉近相同的样本，推远不同的样本，以实现对样本的有效识别，完成行人重识别。与特征表示学习不同，优化度量学习使网络能够聚焦于更多有用的信息，使网络向目标方向优化训练，进一步提升所学习到特征的判别能力和模型的区分能力。Zhao 等人[47]设计一种基于跨模态结构的硬挖掘五元组损失函数，利用硬挖掘采

样方法，在特征嵌入后获得最难的五元组对来约束模态内和跨模态变化，保证模型的收敛性。Ye 等人[48]提出一种双向中心约束排序损失，将交叉模态和模态内约束结合，以应对图像间的模态内差异和跨模态间变化。Liu等人[49]提出一种由跨模态三元组损失和模态内三元组损失组成的双模态三元组损失，用来监督特征学习的目标，通过关注异质数据源差异和模态内同一行人的变化来强化图像中鉴别性信息的学习。Liu 等人[50]提出了一种记忆增强单向度量（Memory-Augmented Unidirectional Metric, MAUM）学习方法，该方法在两个单一方向上学习显式的跨模态度量，并通过基于记忆的增强进一步强化学习到的单向度量，以抑制模态差异，进一步提高了交叉模态的识别准确率。

上述基于特征学习和度量学习的方法在面对复杂的数据结构可能无法有效的提取有用特征，限制模型性能。在模态转换网络方面，随着生成对抗网络（Generative Adversarial Network, GAN）的发展，越来越多人应用 GAN 技术或图片风格转化的方式来解决跨模态 Re-ID 问题。此类方法能够通过模态转换获得多样性的数据样本，强化模型的特征学习，以提高模型的预测能力。如 Wang 等人[51]使用 GAN 来解决跨模态差异问题，通过像素对齐模块将可见光图像合成虚拟的红外图像再与真实的红外图像进行匹配，以缓解图像的模态差异；同时考虑到行人身份信息的完整，利用联合鉴别模块优化图像和特征对之间的分布，确保网络不受新噪声的干扰。Kniaz 等人[52]提出了一个 ThermalGAN（Thermal Generative Adversarial Network）网络框架，利用生成对抗网络将单一的彩色探针图像转换为红外图像的探针集，在红外模式下进行图像匹配，将跨模态检索任务转化为单模态任务。Li 等人[53]采用轻量级生成器生成 X 模态图像，该模态作为辅助与其他两个模态图像共同输入到权值共享的学习器中进行交叉模态学习，达到减少模态差异的目的。Zhang 等人[54]提出了一个中间模态网络（Middle Modality Network, MMN），利用非线性的中间模态发生器将 RGB 和 IR 图像有效的投射到统一的图像空间，生成中间模态图像来辅助网络减少模态差异，最后采用分布一致性损失使生成的图像分布尽可能一致。Liu等人[55]建立一种新的频谱感知特征增强网络（Spectrum-aware Feature Augmentation Network, SFANet），利用信道转换生成跨频谱灰度图像以取代 RGB 图像，经过信道扩展后，将灰度模式和红外模式的图像输入可共享的双路径信息保存网络，前者用

于捕捉特定于模态的低级特征，后者旨在学习灰度和红外模态的共同特征。

1.3 论文的主要研究内容

由于昼夜光照条件的不同，在实际监控中摄像头需要在光源不充足时由可见光模式自动切换为红外模式拍摄行人图像，为实现智能监控的全天运作，论文主要针对 RGB-IR 图像的跨模态行人重识别技术展开研究。通过分析跨模态行人重识别研究的难题，论文分别从弱化由拍摄角度、行人姿态差异、背景杂波、遮挡等引起的图像模态内差异以及增强网络应对跨模态图像间差距的鲁棒性两方面入手，提出了基于深度信息融合的跨模态行人重识别方法研究，其主要研究内容如下：

（1）深度信息融合变分细化蒸馏的跨模态行人重识别

基于不同粒度的全局信息融合策略，提出一种深度信息融合变分细化蒸馏的跨模态行人重识别改进方法。首先，采用全局特征学习方法，在骨干网络后利用多分支和广义均值池化提取图像的全局和细节语义信息，同时将不同分支的信息进行再融合，以避免关键细节信息的遗漏；其次，为实现特征信息的最优表示，基于信息瓶颈理论知识，利用变分细化蒸馏策略将多分支捕获的特征信息进行压缩，抑制背景杂波等噪声对网络的干扰，并保留重要的任务相关信息；最后对不同部分实施不同损失进行监督完成网络的联合优化。

（2）基于多样化深度信息融合嵌入的跨模态行人重识别

基于多样化深度特征信息融合与拓展策略，提出一种多样化深度信息融合嵌入的跨模态行人重识别改进方法，通过加强特征空间嵌入和度量学习策略来优化网络。首先，利用 ResNet50 网络作为骨干提取深度特征；其次，考虑到样本特征的有限性，通过强化特征空间的方式在骨干网络第四层后引入多样化嵌入拓展模块，利用多分支卷积结构生成多样化深度特征信息，并与特征空间中原始的特征信息融合，学习不同的深度信息表示来保证网络训练数据的多样性；最后，使用异质中心共享三元组损失完成类内、类间的特征距离约束，以及增强网络学习跨模态共享特征的能力，联合交叉熵损失提升分类的准确度。

（3）融合多尺度特征与混淆学习的跨模态行人重识别

基于多尺度特征和模态混淆学习策略，提出一种融合多尺度特征与混淆学

习的跨模态行人重识别改进方法。首先，采用全局和局部特征结合的方式分别学习行人的局部细化特征与全局粗糙特征，挖掘不同尺度的语义信息；其次，为充分利用行人间的跨模态共享信息，将模态标签信息与特征信息融合，把深层网络设置为忽略模态信息的结构，通过模糊网络的模态识别反馈，挖掘稳定且有效的模态无关属性应对跨模态差异。最后根据网络应对跨模态及类内变化的需求，采用不同的损失联合优化网络。

1.4 论文的组织结构

全文组织结构共分为六章，具体如下：

第一章绪论部分。从行人重识别技术的研究背景、意义、待解决的难题和智能监控系统流程展开介绍。其次概述了单模态和跨模态行人重识别领域的国内外研究现状。最后对论文的研究内容进行了说明。

第二章对基于深度学习的行人重识别相关技术展开介绍。具体包括卷积神经网络的重要组成、跨模态行人重识别的算法流程、骨干网络结构以及其公开的两个数据集和评价指标。

第三章是深度信息融合变分细化蒸馏的跨模态行人重识别方法研究。首先概述了算法的整理框架；其次，阐述了算法中的各个模块，包括特征提取部分、双重信息聚合模块、变分细化蒸馏策略以及损失函数。然后，从与其他方法对比、与基线精度对比实验说明该算法的先进性，从整体消融实验、不同池化方式对网络的影响以及行人检索排序效果可视化实验分析该算法的有效性。最后对该章内容进行总结。

第四章是基于多样化深度信息融合嵌入的跨模态行人重识别方法研究。首先阐述了该方法的研究思路和算法流程；然后分别介绍了深度特征提取模块、多样化特征嵌入模块和异质中心共享三元组损失函数。最后在公开数据集上通过一系列实验对本章算法进行分析，并对该章节进行总结。

第五章是融合多尺度特征与混淆学习的跨模态行人重识别。首先对整体算法流程和各个模块进行详细介绍；然后，进行实验过程与分析，从与其他先进方法进行比较、精度提升、消融以及行人排序可视化几个方面展开介绍；最后对该章进行总结。

第六章是总结展望，概述论文研究内容和工作，同时分析当前研究存在的不足，并指出了今后的工作方向。

第 2 章 基于深度学习的行人重识别相关技术分析

2.1 引言

随着计算机硬件性能的不断提升，深度学习技术迅速发展。在行人重识别领域，过去依赖手工特征提取的方法已被基于卷积神经网络的特征提取方法取代，使智能化监控系统性能获得极大提升。基于深度神经网络在特征学习方面显著的优势和智能监控场景需求的增加，以深度学习为基础的 Re-ID 技术被普及并得到广泛的认可。

本章是基于深度学习的跨模态行人重识别方法研究，为更好地理解本论文的研究内容，本章从卷积神经网络的四个组成部分、跨模态行人重识别方法流程、用于特征学习的骨干网络结构，以及该技术的两个公共数据集和评价指标展开介绍。

2.2 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）[56]是一种深度学习模型，能够自主学习视频、图像中的有用信息，主要通过权值共享、局部感受野和池化操作实现分类、检测等任务，是现代深度学习领域的重要组成部分。如图 2-1 所示，CNN 网络主要由输入层、卷积层、激活函数、池化层和全连接层五个部分组成。

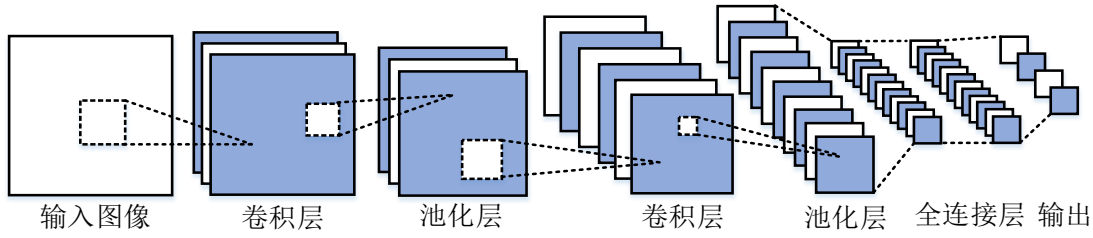


图 2-1 CNN 结构图

2.2.1 卷积层

卷积层是 CNN 的基础，它包括一系列卷积核堆叠的滤波器，利用卷积核对

图像进行特征抽取、特征映射，计算输出更高级的特征。在训练中，网络根据任务要求利用反向传播优化每个卷积单元的卷积尺寸、步长等参数，以实现从图像像素数值组中高效提取不同区域的特征。卷积层操作的实质是对输入信号数据按元素相乘并累加得到卷积结果，其计算过程如图 2-2 所示，以 1 通道 5*5 的输入图像为例，使用 3*3 的卷积核，按照规定步长 1 从左到右，从上到下进行滑动计算输出 3*3 的特征图。而实际的 CNN 中，卷积层由多个卷积核组成且多个卷积核共同参与卷积运算，由公式（2-1）可表述卷积操作。其中，a^{l-1}为输入特征图，k^l 表示卷积核，l 表示第 l 个卷积核，a^l 表示卷积操作后的特征图。对于输入为三通道的可见光图像进行卷积操作，则需要将每个通道与卷积核进行计算，然后对应位置求和，输出特征映射的矩阵表示。

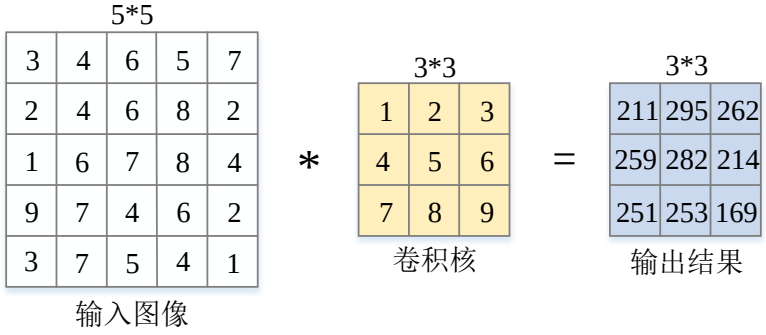


图 2-2 卷积计算过程

$$a_j^l = f\left(b_j^l + \sum_{i \in M_j^l} a_i^{l-1} * k_{ij}^l\right) \tag{2-1}$$

与普通网络相比，CNN 以权重共享和局部连接的方式来控制网络的参数量，提高模型训练的收敛速度。权重共享即利用权内系数相同的滤波器进行卷积处理输入图像，以高效处理图像数据。且卷积层由许多可学习的卷积核集合组成，当处理复杂的图像数据时，卷积核内的神经元无法与上层的所有神经元进行连接，故将神经元与输入数据的局部区域连接，被覆盖的区域称为感受野。感受野大小决定对数据信息理解程度，较大的感受野能够增强对上下文信息的感知能力，较小感受野更关注细节信息。如公式（2-2）所示，每一层的抽象层次可以由感受野的值来推测。

$$l_k = l_{k-1} + \left((f_k - 1) * \prod_{i=1}^{k-1} s_i\right) \tag{2-2}$$

其中，l_{k-1}为前一层的感受野，f_k为第 n 层卷积核的尺寸，s_i为前一层的步幅，l_k为该层感受野尺寸大小。

2.2.2 激活层

卷积操作是通过对输入矩阵和卷积核矩阵进行线性组合来实现的，只能拟合一定的线性模型，因此需要在神经网络中利用激活函数实现非线性，保留并映射卷积层输出的特征，有助于梯度在网络中的反向传播，提升图像数据的表达能力。通常采用的激活函数有 ReLU（Rectified Linear Unit）[57]函数和 Leaky ReLU 函数。

ReLU 函数如图 2-3 的（a）所示，使其部分神经元输出为 0，正值不变直接输出，具有单侧抑制能力，优化了输入值较大时梯度消失的状况，保证网络的稀疏性，减少网络间参数相互依赖，以应对深度神经网络训练过拟合情况。但在输入小于零时梯度为零会导致神经元坏死，即神经元不会被任何数据激活。ReLU 函数如公式（2-3）所示。

$$\mathrm{ReLU}=\max(0,x) \tag{2-3}$$

Leaky ReLU 函数如图 2-3 的（b）所示，该函数保留了 ReLU 函数的优点，在此基础上利用较小的值初始化神经元，修正了 ReLU 函数输入数据为负数神经元坏死的情况，该函数公式如（2-4）所示。

$$\text{Leaky ReLU}=\max(ax,x) \tag{2-4}$$

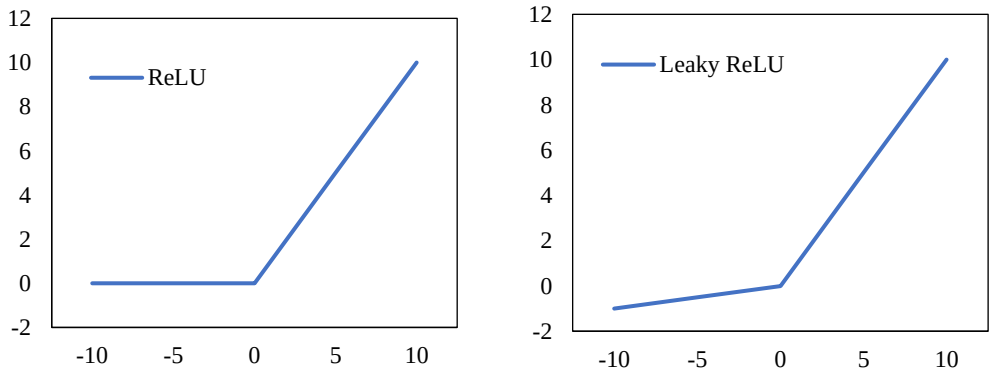


图 2-3 激活函数图

2.2.3 池化层

在 CNN 网络中通常在相邻的卷积层之间嵌入池化层对数据压缩，进行特征选择，保留图片中的重要信息，删除图中与任务不相关的信息，同时降低网络中的参数量，避免过拟合。最大池化（max pooling）和平均池化（avg pooling）为常用的池化方式。其中，最大池化取局部区域中值最大的点；平均池化对局部区域像素取平均值。卷积核大小为 2×2，步长为 2 的池化操作过程如图 2-4 所示。

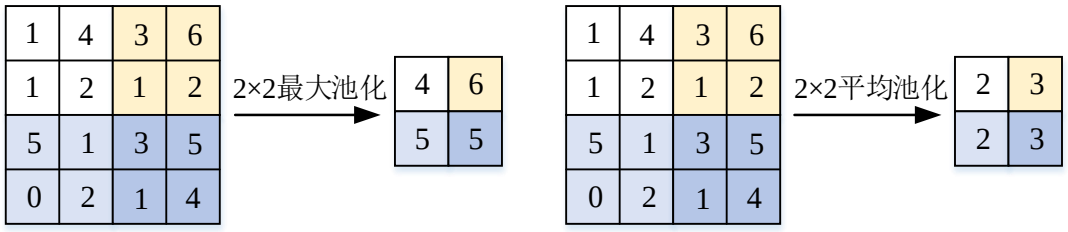


图 2-4 池化操作过程

最大池化可以选择性地保留输入图像的特征，学习突出的特征和纹理信息，摒弃冗余信息。平均池化操作则挖掘图像的背景和全部特征信息，保证信息完整度的同时帮助网络降低图像的空间维度。

2.2.4 全连接层

全连接（Fully Connected, FC）层在卷积神经网络的最后，用来预测图像的类别以实现分类。特征图经过卷积层、激活层和池化层的特征学习和降维后，为避免 CNN 输出时重要特征信息的丢失，利用 FC 把特征映射到样本的标签空间，将池化后的特征矩阵转化为一维特征向量，对数据进行再次降维以实现分类任务。

如图 2-5（a）所描述，CNN 中通常由两个 FC 层组成。第一层将提取到的所有特征展开按照顺序与同一向量空间的所有神经元相连接，确保每个神经元可以接收上层的信息。第二层将具有鉴别性的特征进行线性变换映射到样本标签空间，对图像特征信息进行分类。此外，如图 2-5（b）所示，为了有效缓解邻近层间不同神经元的依赖性，利用 Dorpout 正则化方法随机切断正向传播中的神经元信号的接收和传递。该方法可看作对同一组输入数据每次使用不同的模型进行训练优化，有效增强了模型的鲁棒性，同时避免了过拟合情况。

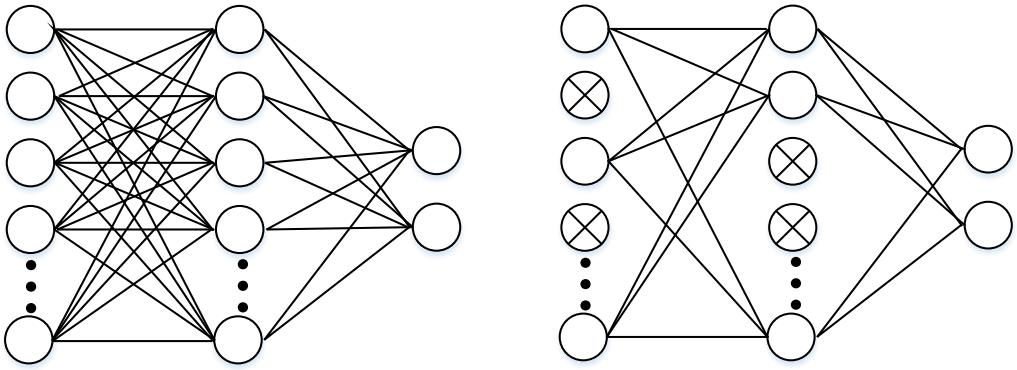


图 2-5 全连接层。（a）未正则化处理的全连接层。（b）Dorpout 正则化后的全连接层。

2.3 跨模态行人重识别流程及基础知识

2.3.1 跨模态行人重识别算法流程

跨模态行人重识别研究旨在不同模态下跨越多个摄像机寻找同一个行人图像。首先，利用能够由可见光模式转换为红外模式的摄像机分别捕捉行人的 RGB 和 IR 图像，并对获取的图像进行预处理构建跨模态行人重识别数据集。其次，以 ResNet50 为骨干网络搭建跨模态 Re-ID 模型，利用双流网络对输入的 RGB 和 IR 行人图片数据进行特征学习。最后对提取的两种模态特征进行度量学习，实现跨模态的查询行人匹配。跨模态行人重识别流程如图 2-6 所示，具体分为以下三部分。

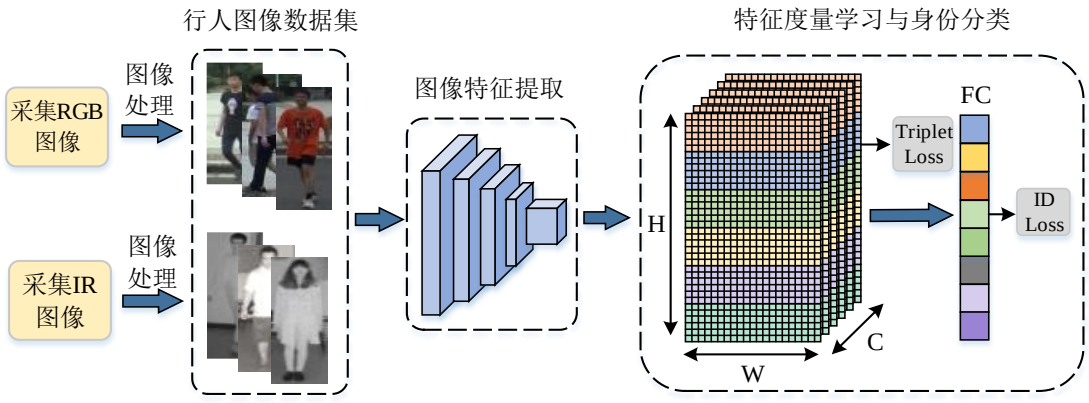


图 2-6 跨模态行人重识别流程图

（1）构建行人图像数据集：对所拍摄的 RGB 和 IR 行人图片和视频进行处理，通过目标检测算法对行人图像进行裁剪和标签处理后，整理跨模态 Re-ID 的图库和待查询图像得到数据集。

（2）图像特征提取：特征学习网络由双流的 ResNet50 为骨干，为进一步弱化模态差异将骨干网络的后几层参数权值共享。然后将行人的 RGB 和 IR 图像输入到具有初始化网络参数的模型进行初步训练，根据模型训练反馈及测试结果优化网络，最终通过反向传播更新、训练并优化网络直至训练出最佳模型。

（3）特征度量学习与身份分类：在特征距离度量学习阶段，通常使用身份损失、三元组损失等多种损失函数来处理网络提取的特征。这些损失通过计算目标图像与图库图像之间的样本特征距离进行排序，以缩小相同身份样本间的距离扩大不同身份的类间差异，或将不同模态的同一身份的特征视为一类以实现行人的身份分类。

2.3.2 ResNet50 网络结构

卷积神经网络通过堆加网络层数来提升图像中复杂特征信息的表示能力。理论上，网络层数越深越能更好的学习图像中的细腻特征，以获得更强大的特征信息增强语义表达能力。但相较于浅层网络，当网络层数足够深时，模型的学习能力反而越来越差，甚至不如较浅层的神经网络。其原因是网络的梯度消失与梯度爆炸，使网络层数越多越难收敛。而残差神经网络（Deep Residual Network, ResNet）通过残差学习的方式在一定程度上缓解了上述现象，在较深层的网络中同样可以保持训练的速度和精度。

残差学习通过跨越连接实现，即在残差块之间添加一个直连通道，使每个残差块的输出可跨越一个或多个网络层至残差块的底部，直接将浅层网络的特征信息与深层特征信息进行叠加，避免了网络的梯度弥散，同时保证梯度传递的有效性，进一步提高了网络的特征表达能力。ResNet 中的残差块结构如图 2-7 所示，首先利用卷积层学习特征，采用 ReLU 函数实现非线性，再通过卷积层进一步提取特征。然后将输入数据与第二个卷积层的输出进行跨层连接来实现残差学习，最后通过 ReLU 函数进行非线性变换输出。

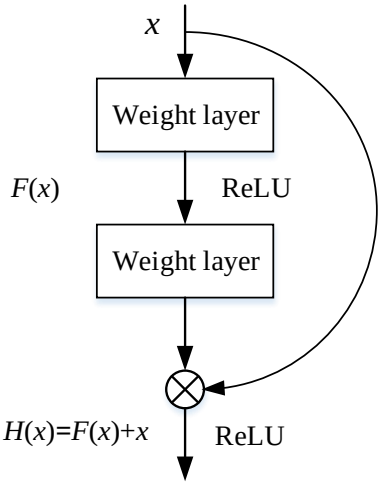


图 2-7 残差结构图

当前大量的跨模态行人重识别工作均基于 ResNet50 骨干上进行改进。如表 2-1所示，ResNet50是多个残差块的堆叠，可分为Conv1，Conv2_x，Conv3_x，Conv4_x，Conv5_x五个阶段。首先，在 Conv1 阶段对形状为（3,244,244）的输入进行预处理，分别使用 7×7，stride 为 2 的卷积层捕获全局特征、BN 层对输入数据进行标准化以及 ReLU 激活函数引入非线性，接着使用 3×3 的最大池化在特征图的宽度和高度上进行降维，得到形状为（64,56,56）的输出。对于后四个阶段，Conv2_x 由三个残差块组成，Conv3_x，Conv4_x，Conv5_x 与 Conv2_x 结构相似，分别包含 4、6、3 个残差块，每个模块都包含三次卷积操作，分别负责对上一阶段的特征进行提取和加深。网络最后采用平均池化和维度为 1000 的全连接层来实现图像分类任务。

表 2-1 ResNet50 网络结构

Layer name	Output size	ResNet50-layer
Conv1	112×112	7×7,64,stride 2
Conv2_x	56×56	3×3max pool,strie 2
		$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$

Conv4_x	14×14	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$
	1×1	Average pool,1000-d fc, softmax

2.4 数据集和评价指标

2.4.1 数据集

颜色信息在大部分行人重识别工作中被视为目标行人查询的重要线索，但对于跨模态行人重识别任务，不能利用颜色信息进行查询匹配。同时单模态行人重识别由于姿态、视角、遮挡等引起的类内变化也是跨模态行人重识别研究的难题之一。为此，Wu[39]等人使用 6 台摄像机收集了第一个跨模态行人重识别领域的大型数据集 SYSU-MM01。如表 2-2 所示，该数据集由 4 个 RGB 摄像机和 2 个 IR 摄像机拍摄的 491 个行人组成，共 287628 张 RGB 图像和 15792 张 IR 图像。根据摄像机的采集环境不同，数据集中的图像被分为室外和室内图像。室外图像由 cam4、cam5 和 cam6 相机在室外拍摄，室内图像由 cam1、cam2 和 cam3 相机在室内环境拍摄。训练时采用 395 个身份的行人图像，其中 RGB 和 IR 图像分别为 22258 张和 11909 张。测试时包含 96 个身份的行人图像，其中探针图像为测试集中的 3803 张 IR 图像，图库集为 301 张 RGB 图像。

测试时设置了全搜索（All-search）和室内搜索（Indoor-search）两种模式。在全搜索模式下检索库中的 RGB 图像为 cam1、cam2、cam4、cam5 相机拍摄，查询库的 IR 图像为 cam4、cam5 相机拍摄。室内搜索模式则仅使用室内图像进行测试，其中检索库为相机 cam1、cam2 拍摄，查询库为 cam3、cam6 相机拍摄。

表 2-2 SYSU-MM01 数据集

摄像头	坐标	室内/室外	光照	行人类别数	RGB 类别数	IR 类别数
1	房间 1	室内	日光	259	400+	-
2	房间 2	室内	日光	259	400+	-

3	房间 2	室内	暗光	486	-	20+
4	大门	室外	日光	493	20	-
5	花园	室外	日光	502	20	-
6	马路	室外	暗光	299	-	20+

RegDB[58]数据集包含 412 个行人的图像数据，分别由一个 RGB 相机和一个热成像相机进行采集，收集了每个行人的 10 张 RGB 和 IR 图像。总计 8240 张图像。训练时，任意选择 206 个行人身份图像，剩下 206 个行人身份图像用于测试。测试阶段，分为可见光到红外（Visible to Infrared）和红外到可见光（Infrared to Visible）两种模式，为保证实验数据的稳定性，进行 10 次交叉验证，最终取平均值作为结果。

2.4.2 评价指标

在实验中，采用标准的累计匹配特征（Cumulative Matching Characteristics, CMC）曲线、平均精度（mean Average Precision, mAP）和平均逆负惩罚（mean of Inverse Negative Penalty, mINP）作为评估标准。CMC 根据相似度计算前 n 个结果中正确检索到图片的百分比，称为 R-n（Rank-n）。其中常采用 Rank-1、Rank-10 和 Rank-20 作为评价指标衡量检索任务的准确度。Rank-n 的值越大表示图像匹配的准确度越高，算法鲁棒性越强。mAP 表示测量所有类之间的平均检索性能，计算公式如 2-5 所示。首先根据各类别检索样本的置信度进行排序，通过计算不同置信度下的精度和召回率来获得每个类别的平均精度（Average Precisin, AP），接着对所有类别的 AP 取均值最终得到 mAP。

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{C} \tag{2-5}$$

其中，AP_i 表示每个类别的平均精度，C 表示类别数。mAP 作为重要的评估指标，其值越大说明模型的效果越好。

mINP 为所有查询样本的平均逆置负样本惩罚率，表示检索最难正确匹配的工作量。如公式 2-6 所示，该评价指标是对 Rank-n 和 mAP 的补充，通过设计一个负惩罚（Negative Penalty, NP）来衡量检索到最难匹配样本的成本。

$$NP_i = \frac{R_i^{hard} - |G_i|}{R_I^{hard}} \tag{2-6}$$

2.5 本章小结

本章对基于深度学习的行人重识别技术分为以下几方面进行简要阐述。首先详细概述了卷积神经网络的主要组成部分，分别为卷积层、激活函数、池化层和全连接层。其次从构建行人图像数据集、图像特征提取以及度量学习三方面阐述了跨模态 Re-ID 的流程，并对其中进行图像特征学习的 ResNet50 结构进行分析。最后对跨模态 Re-ID 的两个公开数据集、测试模式，以及三种评价指标 Rank-*n*、mAP、mINP 进行了介绍。

第 3 章 深度信息融合变分细化蒸馏的跨模态行人重识别

3.1 引言

行人重识别（Person Re-Identification, Re-ID）普遍应用于智能视频监控、智能安防、智慧交通等多个领域，对预防犯罪、维护公共安全、帮助公安刑事侦查、管理城市交通等方面有重要作用。大部分的 Re-ID 研究是在白天或光线充足的场景进行可见光到可见光单模态下的图像搜索[59]。但在实际的智能监控中，为保证监控的连续性，摄像机在夜间通常自动切换到红外模式，故提出跨模态行人重识别研究[60]。昼夜环境的变化使相机拍摄的图像呈现多模态特性，影响了行人特征信息的提取与识别。同时，由于多个非交叠的摄像头在不同位置、距离和角度下进行拍摄，导致所捕获的图像存在巨大的视角差异，进一步增加了行人检索匹配的复杂度。因此，在多模态变化以及不同视角下重要特征信息的挖掘与充分利用问题成为智能监控系统中重要的挑战之一。

为解决以上问题，常见的跨模态行人重识别研究设法利用注意力机制感知重要性权重来优化关键信息的捕获，同时结合全局特征学习具有区分性且模态共享的特征帮助网络实现多模态行人图像的鉴别与分类。如 Ye 等人[61]通过设计 AGW 网络用于关注全局特征，利用深层骨干网络权重共享设置分别学习特定于模态和模态共享的特征表示，并采用注意力机制进行特征加权防止重要语义信息丢失。但 AGW 网络仅利用单一的粗粒度全局共享特征来处理模态差异，缺乏对全局特征中细节信息的考量，以致行人表征的判别性不足；并且采集的特征信息易受背景杂波等噪声的干扰，使网络提取的特征鲁棒性不高，从而影响网络的识别精度。

针对上述深度特征信息挖掘不足与冗余信息干扰的问题，本章从充分利用行人间的重要特征信息和弱化噪声干扰的角度出发，提出了一种深度信息融合变分细化蒸馏的跨模态行人重识别改进方法，该方法旨在学习两种模态间粗细粒度的共享全局特征和实现最小化冗余信息的影响。在框架中，包含提取行人深度信息的双重信息聚合模块（Dual Information Aggregation Module, DIAM）和最小化冗余的变分细化蒸馏（Variational Refinement Distillation, VRD）策略。最

后采用多种损失函数联合的方式对网络的不同部分进行监督学习，以提高网络模型的精度。

3.2 算法流程

本章网络框架基于双流的 ResNet50 骨干网络提出适用于跨模态行人重识别的算法，如图3-1所示，主要包括特征提取模块、双重信息聚合模块和变分细化蒸馏策略。双重信息聚合模块（DIAM）对特征提取模块输出的特征图进行深层细化学习，挖掘全局特征中行人更细腻的特征区域，然后通过变分细化蒸馏（VRD）最大化保留已捕获的行人表征，同时减少与任务无关的信息，来实现最优表示。

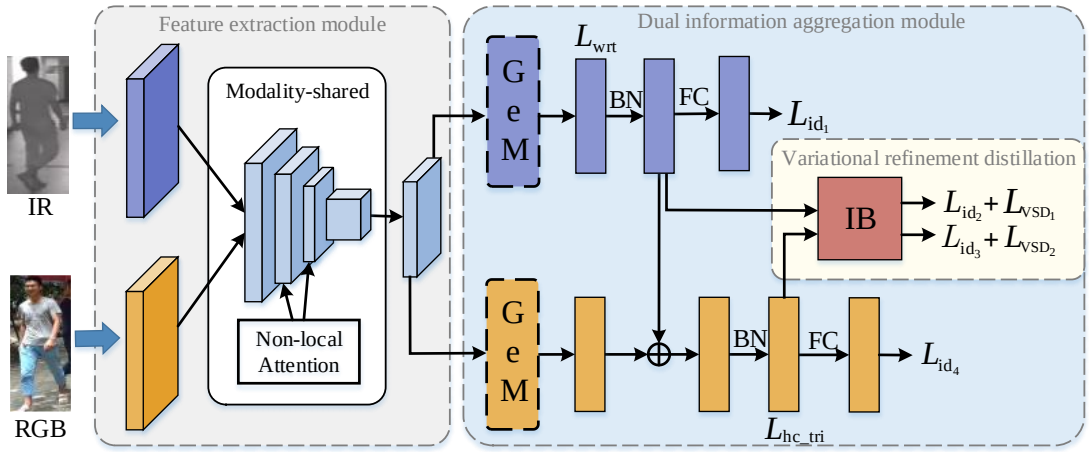


图 3-1 整体网络结构图

3.3 算法原理

3.3.1 特征提取模块

可共享的特征提取模块（Feature extraction module）包括双流的ResNet50网络与注意力模块，如图3-2所示。为缩小模态间差异，将骨干网络的浅层卷积块设置为捕获特定于模态的特征表示；其余四层为模态共享（Modality-shared）特征层，学习公共3D特征空间中来自异构模态的行人共享信息。由于ResNet50网络层数较多，为防止重要的全局语义信息丢失，在共享层的第二层和第三层中分别加入非局部注意力机制（Non-local Attention）[62]，以建立两个较远距离特征像素之间的联系，增强信息共享能力，其原理如公式（3-1）所示。

$$\boldsymbol{Z}_i = \boldsymbol{W}_Z * \phi(\boldsymbol{X}_i) + \boldsymbol{X}_i \tag{3-1}$$

其中，$\boldsymbol{W}_Z$ 为待学习的权重矩阵，$\phi(\bullet)$ 为非局部学习操作，$*$ 表示卷积操作，$\boldsymbol{X}_i$ 表示输入信息，$+\boldsymbol{X}_i$ 表示残差连接。

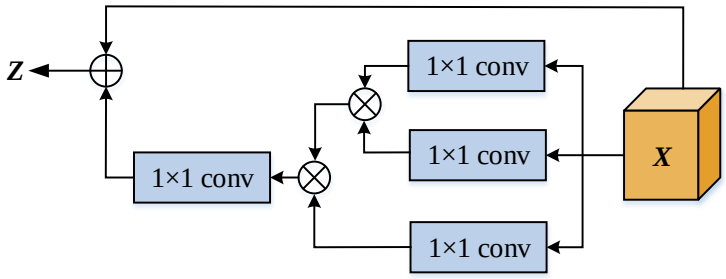


图 3-2 非局部注意力结构图

3.3.2 双重信息聚合模块

在跨模态行人重识别中，行人图像间的模态差异是影响网络性能的关键。为了减少模态差异，捕获图像的全局和关键细节信息（如行人所佩戴的眼镜、帽子及衣物上的图案等），从而获得更有判别性的特征表示就显得尤为重要。因此本章设计了一个双重信息聚合模块，通过构建不同粒度的双分支对骨干网络输出的特征信息进行深层细化学习，来增加特征表示的鉴别能力，如图 3-3 所示。

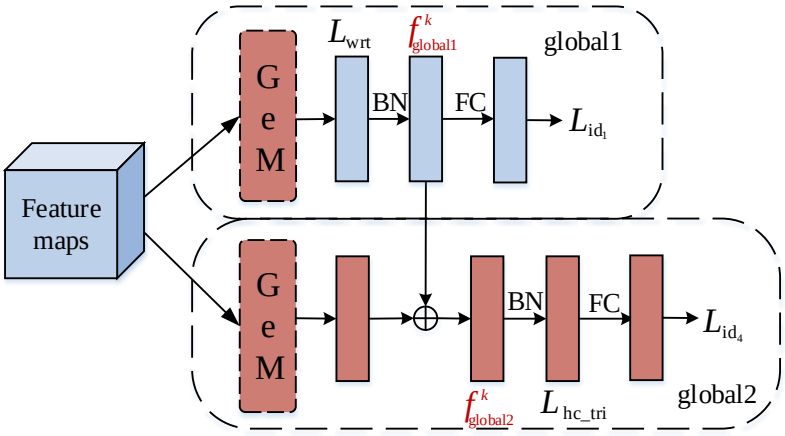


图 3-3 双重信息聚合模块

DIAM 侧重于学习具有类间识别能力和类内紧凑性的高质量特征。为使网络关注行人更细腻的特征区域，在 DIAM 的双分支中采用可学习的广义均值（Generalized-Mean, GeM）池化[63]代替其他池化方式，作为细粒度检索，通过自适应的聚合特征，以产生更紧凑的图像表示。其计算公式如（3-2）和（3-3）

所示。

$$f^{(g)}=[f_1^{(g)}\cdots f_k^{(g)}\cdots f_K^{(g)}]^{\mathrm{T}} \tag{3-2}$$

$$f_k^{(g)}=\left(\frac{1}{|x_k|}\sum_{x\in x_k}x^{p_k}\right)^{\frac{1}{p_k}} \tag{3-3}$$

其中，x为特征输入，$f^{(g)}$表示对应的池化操作输出，$f_k^{(g)}$表示特征图，P_k表示池化参数，x_k表示特征映射的集合。特征向量最终由每个特征图的一个值构成，其维数为 k。P_k作为在网络反向传播中可学习的参数，当$P_k=1$时，f表示平均池化；P_k趋近无穷大时，f表示最大池化。

对于细粒度全局分支（global1），将 GeM 池化提取的特征输入到批处理归一化（Batch Normalization, BN）层，对特征进行归一化。得到 f_{global1}^k，其公式如（3-4）所示。接着将归一化后的特征送入全连接（Fully Connected, FC）层，将维度从 2048 降至 395，同时在批归一化层前后分别以不同的损失处理不同的特征向量，解决同一嵌入空间识别和度量损失不一致的问题。对于粗粒度全局分支（global2），将 GeM 池化聚合的特征与 global1 中归一化约束后的特征相结合，作为粗粒度检索得到 f_{global2}^k，其公式如（3-5）所示。接着将得到的特征输入到 BN 层进行归一化，最后传递给 FC 层对特征进行分类处理。DIAM 在 global1 和 global2 分支的共同作用下提取不同粒度的全局特征，避免单一分支中关键细节信息的遗漏。

$$f_{\text{global1}}^k=\mathrm{BN}[\mathrm{GeM}(x)] \tag{3-4}$$

$$f_{\text{global2}}^k=f_{\text{global1}}^k+\mathrm{GeM}(x) \tag{3-5}$$

其中，x表示特征输入，f_{global1}^k和f_{global2}^k分别表示 global1 和 global2 中第 k 张特征图的特征向量，$\mathrm{GeM}(\bullet)$表示广义均值池化操作，$\mathrm{BN}(\bullet)$表示批量归一化操作。

3.3.3 变分细化蒸馏策略

为进一步提升网络模型的性能，本章将单模态中基于信息瓶颈理论的变分自蒸馏（Variational Self-Distillation, VSD）[64]损失扩展到跨模态学习，对 DIAM 提取的不同粒度信息进行再处理，设计了变分细化蒸馏（Variational Refinement Distillation, VRD）策略。其中，信息瓶颈（Information Bottleneck, IB）为表示学习提供了一个信息理论原理，即保留所有与预测标签相关的信息，同时最小

化冗余信息。虽然 IB 原理已经得到广泛应用，但其中互信息的准确估计仍然是一个挑战。VRD 策略利用 VSD 损失来拟合互信息而不估计他，以避免复杂的设计；并且通过最大化不同粒度的有用信息，同时弱化噪声的影响，来提升表征的鲁棒性。

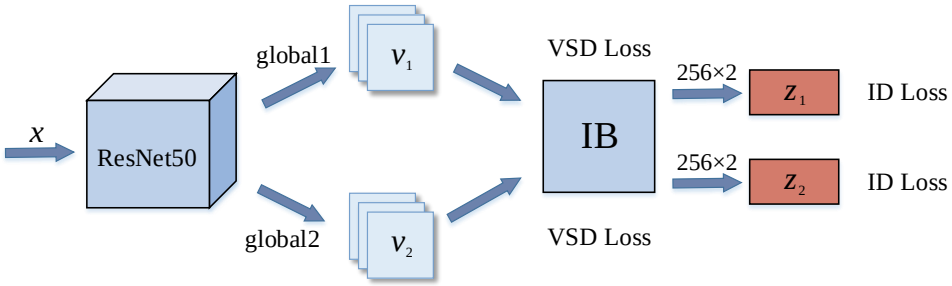


图 3-4 变分细化蒸馏模块

如图 3-4 所示，VRD 策略如下：首先，将两种模态的行人图像 x 输入 ResNet50 中，以学习异构模态行人的共享特征表示。然后利用 global1 和 global2 分支中 BN 层约束后的特征作为输出，以挖掘跨模态行人不同粒度的全局预测信息。接着，把不同粒度的预测信息观察值 v_1 和 v_2 输入一个共享的信息瓶颈（IB）中，结合 VSD 损失最大化保留与任务相关的深层细化信息，同时减少噪声。最终利用交叉熵损失来帮助识别行人身份，进一步增强特征表示判别性。其中，IB 为一个多层感知器，包含两个大小分别为 1024 和 512 的 ReLU（Rectified Linear Units）单元，输出大小为 256×2。变分自蒸馏损失通过保证 IB 产生的表示 z_1 (z_2)最大化的保留 v_1 (v_2)中与标签相关的信息，最小化 v_1 (v_2)中与任务无关的干扰物，来保持表示 z_1 (z_2)的充分性。公式（3-6）和（3-7）分别为表示 z_i 和观测 v_i 的预测信息的分布，其中 i=1,2。

$$P_{z_i} = p(y \mid z_i) \tag{3-6}$$

$$P_{v_i} = p(y \mid v_i) \tag{3-7}$$

x 表示输入数据，y 作为输入标签，v_i 是包含和 x 相同数量的预测信息的观察值，z_i 是信息瓶颈产生的表示。

由于保持充分性的目标等价于最小化 v_i 和 z_i 预测分布之间的差异，故利用公式（3-8），迭代的执行 L_{VSD}，可实现该目标。

$$L_{\mathrm{VSD}_i} = \min_{\theta, f} E_{v_i \sim E_\theta(v_i \mid x)} \left\{ E_{z_i \sim E_\phi(z_i \mid v_i)} \left[D_{\mathrm{KL}} \left(P_{v_i} \,\|\, P_{z_i} \right) \right] \right\} \tag{3-8}$$

其中，i=1,2。θ 和 ϕ 分别表示 ResNet50 和信息瓶颈的参数，$P_{v_i} \,\|\, P_{z_i}$ 表示缩小 P_{v_i}

分布与目标分布 P_{z_i} 之间的距离，D_{KL} 表示 KL 散度，用来衡量两个分布之间的相似性，$D_{\mathrm{KL}}\left(P_{v_i} \| P_{z_i}\right)$操作为将 P_{z_i} 近似为 P_{v_i} 分布，$E[\bullet]$表示对分布求期望，min 表示最小化相应分布期望的相对熵。

3.3.4 损失函数设计

在跨模态行人重识别任务中，为进一步增强行人识别的准确度，本章利用多种损失函数的不同性质共同优化网络模型。对于细粒度全局分支，本章联合加权三元组损失和交叉熵损失对池化后和全连接层后的特征进行处理。对于粗粒度全局特征分支，联合异质中心三元组损失[65]和识别损失对归一化后的特征和全连接层输出的特征分别进行学习。同时将上述双分支得到的特征输入共享的信息瓶颈，利用变分自蒸馏损失和交叉熵损失进行信息蒸馏与分类。因此，网络总损失 L 如（3-9）所示。其中 i=1,2；j=1,2,3,4。

$$L = L_{\mathrm{VSD}_i} + L_{\mathrm{id}_j} + L_{\mathrm{wrt}} + L_{\mathrm{hc\_tri}} \tag{3-9}$$

交叉熵损失用来监督网络学习不随模态而改变的特定信息，提取模态特定的特征，以应对模态内变化，来对行人身份进行分类。L_{id} 的原理公式如（3-10）所示。

$$L_{\mathrm{id}} = -\sum_{i=1}^{K} \ln \frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}} x_i + b_{y_i})}{\sum_{j=1}^{n} \exp(\boldsymbol{W}_j^{\mathrm{T}} x_i + b_j)} \tag{3-10}$$

其中，K 表示一个批次的大小，x_i 表示在第 y_i 类中第 i 个样本的特征，$\boldsymbol{W}_j$ 表示 $\boldsymbol{W}$ 的第 j 行参数，b 表示偏置量。

为帮助网络提高类内相似性和扩大类间差异，应对模态变化，结合加权三元组损失进行监督学习。该损失通过给正样本和负样本进行加权设计，以优化嵌入空间跨模态正负对之间的距离，达到缩小模态差异的目的。L_{wrt} 的原理公式如（3-11）所示。

$$L_{\mathrm{wrt}} = \log\left[1 + \exp\left(\sum_j \boldsymbol{w}_{ij}^{p} d_{ij}^{p} - \sum_k \boldsymbol{w}_{ik}^{n} d_{ik}^{n}\right)\right] \tag{3-11}$$

其中，i、j、k表示每个训练批次中的难三元组。而三元组是由锚点（Anchor）、正样本（Positive）和负样本（Negative）组成。对于锚点 i，P_i 和 N_i 是对应的正样本和负样本，d_{ij}^{p} 和 d_{ik}^{n} 分别表示正样本对的相对距离和负样本对的相对距离，$\boldsymbol{w}_{ij}^{p}$ 和 $\boldsymbol{w}_{ik}^{n}$ 分别表示它们的权重。

为进一步优化深度度量学习，采用异质中心三元组损失进行监督训练，同时处理跨模态差异和模态内变化。该损失不仅提升类内跨模态紧凑性，学习跨模态共享特征，还使用三元组挖掘模态内和模态间的类间可分离性。将其作用在公共特征空间中来指导部分级特征学习，通过样本中心的距离来计算损失。在一个批次中，各模态每个样本的中心距离公式为（3-12）和（3-13）。

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i \tag{3-12}$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{3-13}$$

其中，v_j^i 表示第 i 个行人的第 j 张 RGB 图像，t_j^i 表示第 i 个行人的第 j 张 IR 图像，c_v^i 表示第 i 个行人 RGB 图像的中心，c_t^i 表示第 i 个行人 IR 图像的中心。

对于每个身份，$L_{\text{hc\_tri}}$ 仅关注一个跨模态正对和挖掘在模态内模态间最难的负对。与中心距离公式相结合，$L_{\text{hc\_tri}}$ 的原理公式如（3-14）所示。

$$L_{\text{hc\_tri}} = \sum_{i=1}^{P}\left[mc + \left\|c_v^i - c_t^i\right\|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_v^i - c_n^j\right\|_2\right]_+ + \sum_{i=1}^{P}\left[mc + \left\|c_t^i - c_v^i\right\|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_t^i - c_n^j\right\|_2\right]_+ \tag{3-14}$$

其中，c_v^i 和 c_t^i 表示中心距离来自公式（3-11）和（3-12），c_n^j 表示第 j 个行人的难负样本中心，mc 表示异质中心三元组损失的边缘值。

3.4 实验结果与分析

3.4.1 实验设置

本章实验环境：CPU 为 4 核 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存 16G，显卡 NVIDIA GeForce RTX 2080Ti（显存 11G），操作系统为 Ubuntu 16.04，深度学习框架 pytorch1.1.0。行人图像为 288×144 大小，数据增强通过对图像进行随机裁剪和翻转实现。对于采样策略，SYSU-MM01 中设置了 P=8，K=4，RegDB 中设置了 P=6，K=4，即在 1 个批次中，随机选取 P 个行人身份，每个身份包含 K 张 RGB 图像和 K 张 IR 图像。训练阶段 epoch 大小为 80；初始学习率为 0.1，在 20、50 个 epoch 时学习率衰减 0.1 和 0.01，在前 10 个 epoch 应用热身策略；优化器采用动量为 0.9 的 SGD 进行优化。

3.4.2 实验数据集与评价指标

本章实验采用 SYSU-MM01、RegDB 数据集进行验证，评估指标采用 Rank-*n*、mAP 和 mINP 作为评估标准。

3.4.3 与其他方法进行比较

为了验证本章算法的有效性，将所提方法与以往跨模态行人重识别的先进方法进行比较，主要包括 TONE[40]、D^{2}RL[66]、MAC[67]、Align GAN[51]、Xmodal[53]、FBP-AL[46]、CmCEN[68]、AGW[61]等。实验数据集为 SYSU-MM01 和 RegDB。由对比实验可看出本章方法的优越性，在所有的搜索模式下性能都优于基线网络。其中“—”表示原论文中没有报告的结果。

表 3-1 在 SYSU-MM01 数据集和其他方法对比实验结果（%）

method	**All-search**					**Indoor-search**				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
TONE[40]	12.52	50.72	68.60	14.42	—	20.82	68.86	84.46	26.38	—
HCML[40]	14.32	53.16	69.17	16.16	—	24.52	73.25	86.73	30.08	—
BDTR[41]	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
D^{2}RL[66]	28.90	70.60	82.40	29.20	—	—	—	—	—	—
MAC[67]	33.26	79.04	90.09	36.22	—	33.37	82.49	93.69	44.95	—
DPMBN[69]	37.02	79.46	89.87	40.28	—	44.17	87.12	95.24	54.51	—
Align GAN[51]	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
Hi-CMD[70]	34.94	77.58	—	35.94	—	—	—	—	—	—
LZM[71]	45.00	89.06	95.77	45.94	—	54.28	94.22	98.22	63.92	—
Xmodal[53]	49.92	89.79	95.96	50.73	—	—	—	—	—	—
DSCSN[72]	35.10	77.60	88.90	37.40	—	—	—	—	—	—
mtGAN-D[73]	41.10	82.40	91.90	40.50	—	—	—	—	—	—
FBP-AL[46]	54.14	86.04	93.03	50.20	—	—	—	—	—	—
CmCEN[68]	50.09	86.64	93.29	49.46	37.17	56.84	93.04	97.69	63.55	—
Baseline[61]	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
Ours	**56.36**	**90.29**	**95.69**	**54.96**	**42.01**	**62.20**	**94.26**	**97.74**	**69.11**	**65.50**

表 3-2 在 RegDB 数据集和其他方法对比实验结果（%）

method	**Visible to Infrared**	**Infrared to Visible**

	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
HCML[40]	24.44	47.53	56.78	20.08	—	21.70	45.02	55.58	22.24	—
BDTR[41]	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
D²RL[66]	43.40	66.10	76.30	44.10	—	—	—	—	—	—
MAC[67]	36.43	62.36	71.63	37.03	—	36.20	61.68	70.99	36.63	—
Align GAN[51]	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[53]	62.21	83.13	91.72	60.18	—	—	—	—	—	—
DSCSN[72]	60.80	78.60	85.90	60.00	—	—	—	—	—	
mtGAN-D[73]	65.60	84.20	89.60	60.00	—	65.80	85.80	91.40	59.60	—
FBP-AL[46]	73.98	89.71	93.69	68.24	—	70.05	89.22	93.88	66.61	—
CmCEN[68]	74.03	88.25	92.38	67.52	51.22	74.22	87.22	91.99	67.36	—
Baseline[61]	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
Ours	**78.88**	**91.36**	**94.88**	**70.74**	**55.37**	**76.74**	**89.77**	**93.13**	**68.71**	**51.78**

由表 3-1 和表 3-2 可知，本章算法与 AGW[61]算法相比，在 SYSU-MM01 数据集全搜索模式下 Rank-1、mAP 和 mINP 上分别高出 8.86%、7.31%、6.71%；在 RegDB 数据集可见光图像到红外图像搜索设置下 Rank-1、mAP 和 mINP 分别提升了 8.83%、7.31%、6.71%。由于 AGW 算法采用的网络框架仅关注单一的全局特征，缺乏对网络细节信息的考量以及对冗余信息的处理，导致行人识别率不高。针对以上不足，本章利用不同粒度的双支路来构建网络，提取深层全局和细节特征，增加了对行人间细节信息的关注，并且对提取的全局和细节信息进行变分蒸馏，弱化了冗余信息的影响，从而使网络的识别精度得到提升。与 Xmodal[53]算法相比，本章算法在 SYSU-MM01 数据集全搜索模式下 Rank-1 和 mAP 分别高出 6.44%、4.23%；在 RegDB 数据集可见光图像到红外图像搜索设置下 Rank-1 和 mAP 分别高出 16.67%、10.56%。Xmodal 算法采用 x 模态来辅助跨模态学习，但生成的 x 模态图像中可能包含新的噪声，从而影响网络模型的鲁棒性。而本章通过最大化有用信息，同时减少无关干扰的方法进行特征学习，由对比实验结果可知本章网络的鲁棒性更强。本章算法与 FPB-AL[46]方法相比，在 SYSU-MM01 数据集全搜索模式下 Rank-1 和 mAP 分别高出 2.22%和 4.76%；在 RegDB 数据集可见光图像到红外图像搜索设置下 Rank-1 和 mAP 分别高出 4.9%、2.5%。FPB-AL 与本章算法都通过提取行人的细节信息来提升表征能力，不同的是本章算法在此基础上进行了降噪处理，使模型性能进一步提升。

CmCEN[68]算法与本章在同一基础上进行改进，与其相比，本章算法在 SYSU-MM01 数据集全搜索模式下 Rank-1、mAP 和 mINP 分别高出 6.27%、5.5%、4.84%；在 RegDB 数据集可见光图像到红外图像搜索设置下 Rank-1、mAP 和 mINP 分别高出 4.85%、3.22%、4.15%。CmCEN 利用关键信道取代无用信道，来提取更多的关键特征信息，但缺少本章算法中对行人细腻区域的关注；同时 CmCEN 采用生成对抗模块使跨模态特征分布尽可能一致，但生成的额外模态特征，可能会丢弃原始特征中有用的细节信息或引入新的冗余信息，使模型识别的准确度下降。本章算法利用 VRD 策略，在尽可能不丢失有效信息的同时弱化噪声影响，以增强表征辨别能力。

本章所提方法在未引入额外参数的基础上进行优化改进，与 CmCEN 方法相比本章模型训练的数据复杂度较低。同时在精准度提升的基础上所提方法与基线相比网络训练时间不变，增强了在实际应用中的泛化能力。

3.4.4 实验分析

为了展现本章总体网络框架的有效性，如图 3-5 所示，在 SYSU-MM01 数据集中对比了本章算法改进前后的首张图片命中率 Rank-1 和平均精度 mAP 的变化关系。由图（a）和图（b）可以看出，在训练阶段本章所提网络的 Rank-1 和平均精度都始终位于基线网络上方，最终实验结果也高于基线网络的精度。

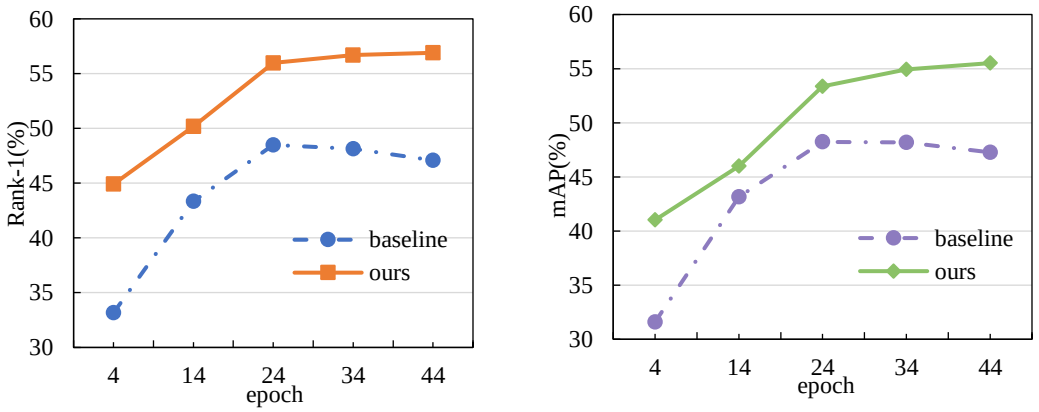


图 3-5 训练阶段不同评价指标变化图。（a） Rank-1 变化过程。（b） mAP 变化过程。

本章中，在 SYSU-MM01 数据集的单镜头全搜索和室内搜索模式下进行了一系列消融实验，来评估所提网络和模块的有效性。

如表 3-3 所示，DIAM 表示将网络框架设置为双重信息聚合结构；VDR1 表示仅对 global1 进行变分蒸馏；VDR2 表示仅对 global2 进行变分蒸馏；VDR 为变分细化蒸馏策略。其中，Ours1 表示在 Baseline 基础上添加 DIAM，由实验结果可知，采用设计的 DIAM 与基线相比模型性能得到明显提升，即在全搜索模式（室内搜索模式）下 Rank-1 和 mAP 分别增加了 6.16%、4.74%（3.47%、2.04%），可见 DIAM 中有效的粗细粒度信息能够被网络学习到。Ours2 表示在 Ours1 基础上对 global1 细粒度全局信息进行变分蒸馏，与 Ours1 方法相比，在全搜索模式（室内搜索模式）下 Rank-1 和 mAP 增加了 1.06%和 1.45%（2.75%和 2.78%）；Ours3 表示在 Ours1 基础上对 global2 粗粒度全局信息进行变分蒸馏，与仅有 DIAM 相比网络精度也有一定提升，即在全搜索模式（室内搜索模式）下 Rank-1 和 mAP 增加了 0.77%、1.08%（1.21%、1.44%）；实验结果说明变分蒸馏策略无论作用在细粒度全局分支还是粗粒度全局分支，都能够削弱图像中噪声的影响。Ours4 表示本章的总体网络框架，即将 DIAM 与 VRD 策略相结合，与仅添加 DIAM 相比，在全搜索模式（室内搜索模式）下 Rank-1 和 mAP 增加了 2.48%和 2.33%（4.36%和 3.94%），同时与 Ours2 和 Ours3 方法相比网络精度也得到明显提升，从而验证了变分细化蒸馏策略的重要性。最终，本章算法与基线相比在全搜索模式（室内搜索模式）下 Rank-1 和 mAP 分别提升了 8.86%、7.31%（8.03%、6.14%），改进算法的有效性得到了验证。

表 3-3 SYSU-MM01 数据集下的消融实验（%）

Method	DIAM	VDR1	VDR2	VRD	All-search			Indoor-search		
					R-1	mAP	mINP	R-1	mAP	mINP
Baseline	×	×	×	×	47.50	47.65	35.30	54.17	62.97	59.23
Ours1	✓	×	×	×	53.66	52.39	39.29	57.64	65.01	61.19
Ours2	✓	✓	×	×	54.72	53.84	41.14	60.39	67.79	64.20
Ours3	✓	×	✓	×	54.43	53.47	40.37	58.85	66.45	62.89
Ours4	✓	×	×	✓	**56.36**	**54.96**	**42.01**	**62.20**	**69.11**	**65.50**

另外，为探索双粒度分支中不同池化方式对网络性能的影响，在 SYSU-MM01 数据集下设计了一系列对比实验，如表 3-4 所示。当网络仅有单分支 global1 时，分别采用广义平均池化和均值池化进行对比。实验结果发现 GeM 池化相较于平均池化性能更好，说明 GeM 池化可捕获特定区域的鉴别特征，进行细粒度检索，故 global1 采用 GeM 池化方法。在此基础上分别对 global2 分支采

用最大池化、平均池化和广义均值池化进行对比实验。实验发现在双分支中都采用广义均值池化时，首张图片击中率和平均精度的值最大，这说明 GeM 池化能够代替最大池化和平均池化，实现自适应的聚合空间信息，提高特征的映射能力。因此，本章采用 GeM 池化来构建不同粒度的双分支。

表 3-4 不同池化方式的对比实验结果（%）

Method		All-search			Indoor-search		
global1	global2	Rank-1	mAP	mINP	Rank-1	mAP	mINP
GeM	—	47.50	47.65	35.30	54.17	62.97	59.23
AVG	—	47.67	46.98	34.30	51.46	59.67	55.81
GeM	MAX	49.83	47.62	33.53	53.98	61.59	57.33
GeM	AVG	52.32	50.61	37.37	57.39	64.61	60.61
GeM	**GeM**	**53.66**	**52.39**	**39.29**	**57.64**	**65.01**	**61.19**

为更直观的展现算法的识别效果，验证所提出的方法的先进性，从公开的大型数据集 SYSU-MM01 中分别挑选两位行人图片，将基线和所提方法在数据集上获得的 Rank-10 检索结果进行可视化，如图 3-6 所示。其中，在 SYSU-MM01 数据集中采用最困难的全搜索模式 Single-hot 的设置下进行检索。行人图像上方的数字为绿色表示检索正确，红色表示检索错误。实验结果表明，与基线方法相比本章所提算法能够有效的弥补网络表征不足从而提升识别精度。

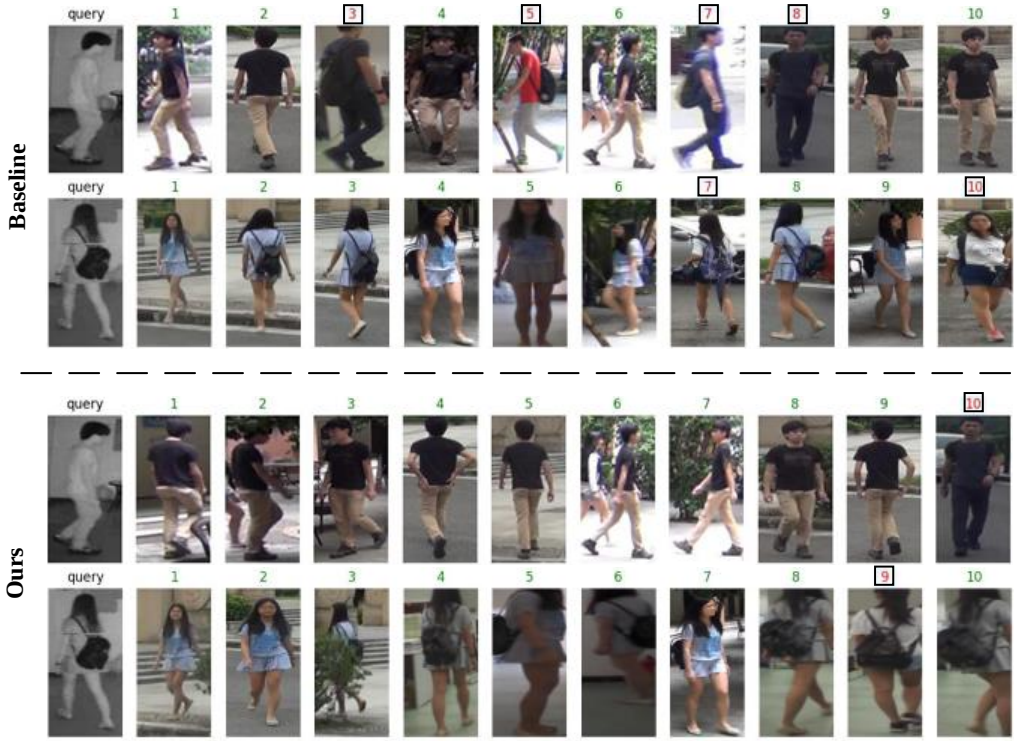


图 3-6 在 SYSU-MM01 数据集上的可视化结果图

3.5 本章小结

本章提出了一种深度信息融合变分细化蒸馏的跨模态行人重识别改进方法。该方法使用双重信息聚合模块和变分细化蒸馏策略，缓解了复杂智能监控场景中由多模态特征变化以及多视角变化引起的重要特征信息挖掘不充分和识别率降低的问题。前者通过提取与融合跨模态不同粒度的深度全局特征，同时学习行人粗糙和细腻的特征区域，增加了对行人间重要细节信息的关注；后者对捕获的信息进行再压缩，实现相关信息最大化和弱化背景杂波等干扰噪声的影响。最终两者相结合在多损失联合约束的作用下，增强行人检索的稳定性和准确度。在两个公共数据集上与目前的先进方法进行对比，通过大量的实验，充分验证了所提方法的有效性。本章算法与基线 AGW 算法相比，在 SYSU-MM01 数据集全搜索设置下 Rank-1、mAP 分别高出 8.86%、7.31%，在 RegDB 数据集 RGB-IR 设置下 Rank-1 和 mAP 上分别提升了 8.83%、4.37%，有效提升了网络识别准确度和模型收敛速度，提高了跨模态行人重识别模型的最终性能。在保障社会安全、开展刑事侦查以及智能交通管理等方面本章改进的跨模态行人重识别方法能够辅助智能监控系统进行人员追踪，实现更高效、准确的搜索。在下一章节中，将继续优化模型，尝试构建更好的深度信息融合网络。

第 4 章 基于多样化深度信息融合嵌入的跨模态行人重识别

4.1 引言

行人重识别作为智能监控系统的关键环节，能够弥补固定摄像头的视觉局限，在无法捕获行人清晰的脸部图像情况下，对行人的身体、四肢、衣服、鞋子等特征进行提取，并识别行人身份。该技术能够在智能监控视频下追踪嫌疑人、寻找失踪人员，对维护公共安全和辅助刑事侦查有重要意义。然而，在实际监控场景中，跨模态行人重识别任务除了面临图像模态间和模态内的差异。还面临由摄像机硬件的局限和复杂环境条件，使图像像素较低、细节、边缘模糊不清等影响图像质量的问题，以至有效的数据训练样本减少，增加了行人检索匹配的难度。

论文第三章算法采用深度信息融合和变分蒸馏策略缓解了多视点匹配、多模态变化导致的模态内和模态间差异，完成跨模态行人重识别中的特征提取任务，并联合交叉熵损失、加权三元组损失和异质中心三元组损失对网络的不同部分进行特征约束。该算法虽然能够较好的改善基线网络的不足，但在复杂的监控环境中，训练样本数据的不充足同样局限了特征学习的能力，从而导致行人匹配准确度不高。同时，当跨模态图像间差异较大时，异质中心三元组损失易受影响，不能很好的监督网络集中于优化模态共享特征。

于是，本章提出一种多样化深度信息融合嵌入的跨模态行人重识别改进方法。针对现有网络模型中有效训练数据不充足以及不同模态图像之间较大模态差异的问题，该算法在特征提取阶段，利用多样化深度特征信息融合来增强特征嵌入空间。具体地说，通过在深层骨干网络中生成更多的共享特征，再与上一层的深度特征信息融合，以丰富训练数据。对于度量学习，利用异质中心共享三元组损失来约束类内、类间的特征距离，同时提取不受模态影响和语义一致性的特征。最后，利用多损失联合优化网络模型，增强算法分类能力，提高识别准确度。

4.2 算法流程

本章提出一种多样化深度信息融合嵌入的跨模态行人重识别改进算法，如图 4-1 所示，该算法流程由以下三个部分组成。

（1）在数据预处理阶段，分别采用通道级随机擦除（Channel-level Random Erasing, CRE）和通道可交换增强（Channel exchangeable Augmentation, CA）对输入的红外和可见光图像进行处理。前者利用模拟随机遮挡来丰富训练样本的多样性；后者为了增强特征表示对颜色变化的适应性，在生成图像时随机调整颜色通道的顺序，以生成与颜色无关的图像。

（2）对于深度特征提取，本章使用 ResNet50 作为骨干网络，在其中嵌入非局部注意力机制学习像素点间长距离相互作用，以提升模型的特征表达能力。并将多样化嵌入拓展模块作用于网络中增强特征嵌入空间，丰富训练数据，提高模型鲁棒性。

（3）利用交叉熵损失和异质中心共享三元组损失共同监督，约束行人的类内类间距离，同时使网络提取不受模态影响、语义一致性特征，改善网络性能。

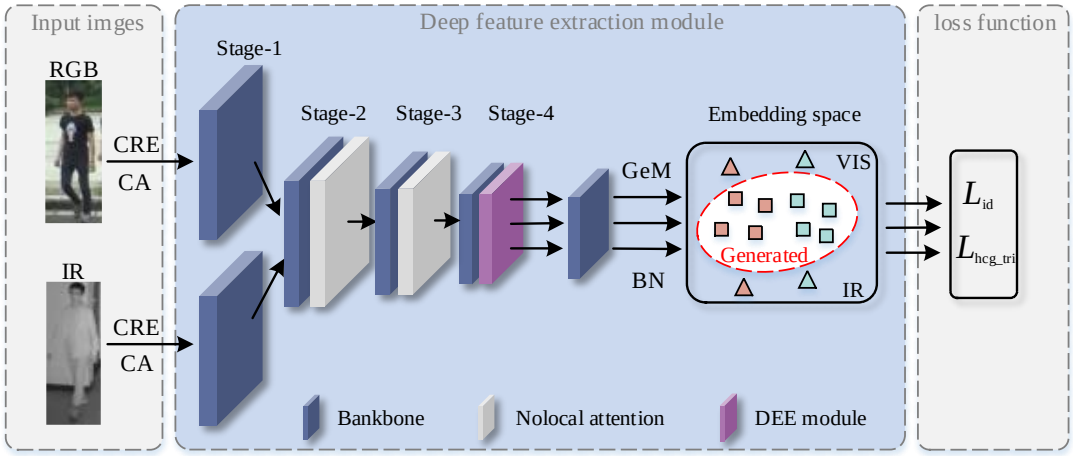


图 4-1 整体网络结构

4.3 算法原理

4.3.1 深度特征提取模块

本章采用 ResNet50 作为骨干网络（Bankbone）来提取行人图片的特征，其中将第一层（Stage-1）设置为权重不共享的双流网络，分别处理来自不同模态图像的特征，而后四层通过网络参数共享来学习模态共享的特征。为防止关键

语义信息的丢失，在骨干网络的第二（Stage-2）、三层（Stage-3）后加入非局部注意力（Nolocal Attention）机制，建立图像中不同像素点之间联系，关注有效信息以增强跨模态共享特征的提取能力。同时为缓解训练数据不足的问题，在骨干网络第四层（Stage-4）后添加多样化嵌入拓展模块，通过增强嵌入空间以生成更多的特征嵌入，来丰富行人的特征描述符。

4.3.2 多样化嵌入拓展模块

跨模态行人重识别任务面临最大的挑战为图像间的模态差异，由于网络模型的训练数据匮乏，导致网络不能有效挖掘不同的跨模态线索来缓解模态差异，从而影响行人身份的正确分类。为有效降低跨模态差异，本章在骨干网络的共享层后加入多样化嵌入拓展[74]（Diversified Embedded Expansion, DEE）模块，通过在嵌入空间中生成不同的特征信息，来强化网络共享特征的嵌入空间，并融合多样化的行人深度特征信息以缓解模态差异，提升算法的鲁棒性。

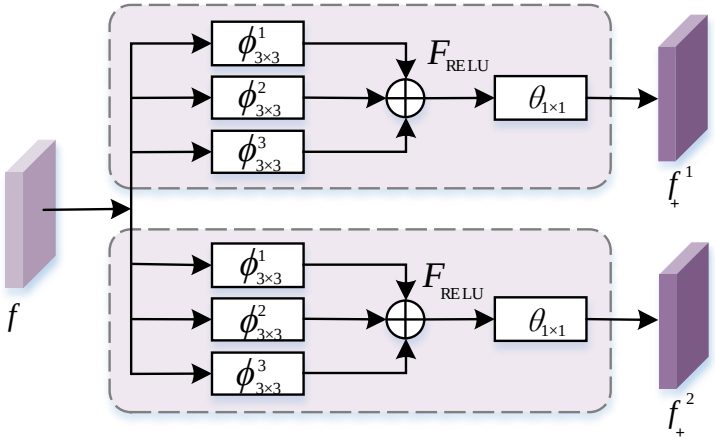


图 4-2 多样化嵌入拓展模块

如图 4-2 所示，DEE 模块主要利用多分支卷积生成更多的特征嵌入来丰富跨模态任务的训练数据。具体地说，对于 DEE 的每个分支，首先将包含红外和可见光特性的特征图输入该模块，使用三个 3×3 的膨胀卷积层对输入的特征图 f 进行处理以扩大感受野，即通过不同的膨胀率（1,2,3）把特征映射的数量减少至其原始特征大小 f 的四分之一大小，再将所获取的特征图拼凑成一个新的特征图。接着采用 ReLU 激活函数实现网络的稀疏性以及提高不同嵌入拓展模块的非线性表示的能力，使模型更好的挖掘相关特征，拟合训练数据。最后，使用一个 1×1 的卷积对得到的特征映射进行维度调整，使其与输入 f 的特征维度保

持一致。因此，可得到第 i 条分支生成的嵌入 f_+^i 为公式（4-1）所示。

$$f_+^i = \theta_{1\times1}\left(F_{\mathrm{RELU}}\left(\varphi_{3\times3}^1(f)+\varphi_{3\times3}^2(f)+\varphi_{3\times3}^3(f)\right)\right) \tag{4-1}$$

其中，i=1,2。$\theta_{1\times1}$ 表示 1×1 的卷积，F_{RELU} 表示 ReLU 激活层，$\varphi_{3\times3}^1$、$\varphi_{3\times3}^2$ 和 $\varphi_{3\times3}^3$ 表示膨胀率分别为（1，2，3）的膨胀卷积。生成的新嵌入与 DEE 模块输入的原始特征图融合，共同作为骨干网络下一阶段的输入，丰富后层语义信息以进一步提高网络模型的鲁棒性。

4.3.3 损失函数

为学习更强泛化能力的特征，增强网络模型在跨模态检索任务中应对行人图像交叉模态和模态内差异的能力，本章使用异质中心共享三元组损失和交叉熵损失联合进行优化训练。

传统三元组损失通过设定一个边界常量，迫使模型中正样本与锚点之间距离更小，负样本与锚点的距离更大，来实现增大类间和缩小类内距离的目的。而加权三元组损失在此基础上对正样本和负样本进行加权设计，以优化嵌入空间跨模态正负对之间的距离，从而实现缩小模态差异的目的。但该损失通过锚与所有其他样本的比较来计算损失，具有强大的约束，对于存在的一些异常值可能过于严苛的限制成对距离，从而形成不利的三联体来破坏其他的成对距离。

针对上述问题，本章采用基于样本中心的三元组损失替换基线网络中的加权三元组损失，使用每个样本的中心作为身份，将基于样本的三元组约束放松到基于样本中心的三元组约束。在训练时利用采样策略，即每次迭代随机采样 P 个行人样本，每个样本分别在可见光和红外模态提取 K 张图片，共 $2P\times K$ 张图片。故，每个批次利用公式（4-2）和（4-3）可以得到 P 个行人的 RGB 图像类别中心 c_v^i 和 P 个行人的 IR 图像类别中心 c_t^i，其中 v_j^i 表示第 i 个行人的第 j 张 RGB 图像，t_j^i 表示第 i 个行人的第 j 张 IR 图像。

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i \tag{4-2}$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{4-3}$$

同时，为进一步帮助网络学习模态共享特征应对跨模态变化，采用不区分

输入图像的模态，仅根据特征间距离进行划分正负样本的思想，提出一种异质中心共享三元组损失，以增强网络的共享能力。将 PK 采样策略和中心距离公式结合，则异质中心共享三元组损失可定义为式（4-4）所示。该损失不仅放松了约束，还保留了在公共特征空间同时处理两个模态的类内类间变化的特性。该特性的核心思想为通过最小化跨模态正样本中心的成对距离拉近相同身份行人间的特征，以及在其他样本中心中挖掘最难的负样本辨别不同身份的特征。

$$\begin{aligned} L_{\text{hcg\_tri}} = &\sum_{i=1}^{P}\left[\rho_1 + \left\|c_v^i - c_t^i\right\|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_v^i - c_n^j\right\|_2\right]_+ + \sum_{i=1}^{P}\left[\rho_1 + \left\|c_t^i - c_v^i\right\|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_t^i - c_n^j\right\|_2\right]_+ \\ &+\sum_{i=1}^{2PK}\left[\max_{\forall y_i = y_j} D\left(f_i, f_j\right) - \min_{\forall y_i \neq y_j} D\left(f_i, f_k\right) + \rho_2\right]_+ \end{aligned} \tag{4-4}$$

其中，c_n^j 表示第 j 个行人的难负样本中心，$D\left(f_i, f_j\right)$ 表示特征嵌入 f_i 和 f_j 之间的距离度量，y_i 表示第 i 个行人的身份标签，ρ_1 和 ρ_2 分别表示不同数值的边缘值。

模态内的变化也是影响跨模态行人重识别任务的关键因素。为确保分类的准确性，本章使用交叉熵损失提取模态的特定特征来应对模态内的变化，并且通过衡量真实概率与预测概率之间的分布差异对行人身份进行分类。将跨模态行人重识别任务视为一个多分类问题，交叉熵损失可表示为公式（4-5）。

$$L_{\text{id}} = -\sum_{i=1}^{K}\ln\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}} x_i + b_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}} x_i + b_j)} \tag{4-5}$$

其中，K 表示一个批次的大小，x_i 表示属于 y_i 第 i 个图片的特征，$\boldsymbol{W}_j$ 表示 $\boldsymbol{W}$ 的第 j 行参数，y_i 表示输入样本特征 x_i 的身份标签，b 表示全连接层的偏置量。

在训练阶段，异质中心共享三元组损失对网络池化后的特征进行优化，交叉熵损失用来监督全连接层输出的特征。故，总的损失函数如公式（4-6）所示。

$$L_{\text{all}} = L_{\text{hcg\_tri}} + L_{\text{id}} \tag{4-6}$$

4.4 实验结果与分析

4.4.1 实验设置

本章实验环境为显卡 NVIDIA GeForce RTX2080Ti（显存 11G）、CPU 为 4 核 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存 16G，操作系统为 Ubuntu

16.04，深度学习框架 pytorch1.1.0。行人的图像大小设定为 288×144，对图像采用随机裁剪和翻转实现数据增强。采样策略中，SYSU-MM01 数据集设置 P=8，K=4，每个训练批次包含 32 张 RGB 图像和 32 张 VIS 图像；RegDB 数据集设置 P=6，K=4，每个训练批次包含 24 张 RGB 图像和 24 张 VIS 图像。其中在 RegDB 数据集下采用文献[74]中的设置，删除 ResNet50 的最后一层并在第三层之后添加 DEE 模块。模型的初始学习率设定为 0.01，前 10 个 epoch 使用热身策略，随后学习率增加至 0.1，再经过 20 次、60 次和 120 次训练后，学习率分别衰减为 0.01、0.001 和 0.0001，共训练 150 个 epoch。训练优化时 SGD 动量参数设置为 0.9。

4.4.2 实验数据集与评价指标

本章实验采用 SYSU-MM01、RegDB 数据集进行验证，在测评时，采用 Rank-n、mAP、mINP 作为评价指标。

4.4.3 与其他方法进行对比

本章将所提方法与现有的跨模态行人重识别方法进行比较，表 4-1、表 4-2 分别为在 SYSU-MM01 数据集和 RegDB 数据集上的实验对比结果，主要包括 DDAG[44]、Xmodal[53]、Align GAN[51]、AGW[61]、cm-SSFT[75]、MCLNet[76]、SFANet[55]、PMT[77]等方法。其中“—”表示原论文中没有报告的结果。

表 4-1 在 SYSU-MM01 数据集和其他方法对比实验结果（%）

method	All-search					Indoor-search				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[39]	14.80	54.12	71.33	15.95	—	20.58	68.38	85.79	26.92	—
TONE[40]	12.52	50.72	68.60	14.42	—	20.82	68.86	84.46	26.38	—
HCML[40]	14.32	53.16	69.17	16.16	—	24.52	73.25	86.73	30.08	—
BDTR[41]	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
eBDTR[48]	27.82	67.34	81.34	28.42	—	—	—	—	—	—
D^{2}RL[66]	28.90	70.60	82.40	29.20	—	—	—	—	—	—
MAC[67]	33.26	79.04	90.09	36.22	—	33.37	82.49	93.69	44.95	—
DPMBN[69]	37.02	79.46	89.87	40.28	—	44.17	87.12	95.24	54.51	—

Align GAN[51]	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
LZM[71]	45.00	89.06	95.77	45.94	—	54.28	94.22	98.22	63.92	—
Xmodal[53]	49.92	89.79	95.96	50.73	—	—	—	—	—	—
DDAG[44]	54.75	90.39	95.81	53.02	—	61.02	94.06	98.41	67.98	—
TSLFN[43]	59.96	91.50	96.82	54.95	—	59.74	92.07	96.22	64.91	—
AGW[61]	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
cm-SSFT[75]	61.60	89.20	93.90	63.20	—	70.50	94.90	97.70	72.60	—
MCLNet[76]	65.40	93.33	97.14	61.98	47.39	72.56	96.98	99.20	76.58	72.10
SFANet[55]	65.74	92.98	97.05	60.83	—	71.60	96.60	99.45	80.05	—
PMT[77]	67.09	94.56	—	64.25	50.89	—	—	—	—	—
Ours	**69.98**	**96.29**	**98.80**	**66.71**	**52.83**	**77.38**	**96.85**	**99.17**	**80.75**	**77.13**

表 4-2 在 RegDB 数据集和其他方法对比实验结果（%）

method	**Visible to Infrared**					**Infrared to Visible**				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[39]	17.75	34.21	44.35	18.90	—	16.63	34.68	44.25	17.82	—
HCML[40]	24.44	47.53	56.78	20.08	—	21.70	45.02	55.58	22.24	—
BDTR[41]	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
eBDTR[48]	34.62	58.96	68.72	33.46		34.21	58.74	68.64	32.49	
MAC[67]	36.43	62.36	71.63	37.03	—	36.20	61.68	70.99	36.63	—
Align GAN[51]	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[53]	62.21	83.13	91.72	60.18	—	—	—	—	—	—
AGW[61]	70.05	—	—	66.37	50.19	70.49	87.12	91.84	65.90	51.24
cm-SSFT[75]	72.30	—	—	72.90	—	71.00	—	—	71.70	—
MCLNet[76]	80.31	92.70	96.03	73.07	57.39	75.93	90.93	94.59	69.49	52.63
SFANet[55]	76.31	91.02	94.27	68.00	—	70.15	85.24	89.27	63.77	—
Ours	**85.38**	**95.26**	**97.42**	**78.44**	**63.04**	**87.08**	**96.27**	**98.03**	**79.79**	**65.76**

根据表 4-1 和表 4-2 的数据，本章方法在以下数据集的三种评价指标显示出明显的优势，超越了许多其他方法。其中，改进后的本章方法在 SYSU-MM01 数据集的全搜索模式下 R-1 达到 69.98%、室内搜索模式下 R-1 达到 77.38%，mAP 分别达到 66.71%、80.75%，mINP 分别达到 52.83%、77.13%；在 RegDB 数据集的两个模式下 R-1 分别达到 85.38%、87.08%，mAP 分别达到 78.44%、79.79%，mINP 分别达到 63.04%、65.76%，表现出良好的性能。相对于使用度

量学习的 BDTR[41]、DDAG[44]等方法，以及将对抗生成网络和度量学习相结合的 D^2RL[66]、Align GAN[51]和 Xmodal[53]等方法，本章具有更高的网络精度，说明本章方法比上述度量学习方法及对抗生成策略具有更强的有效性。

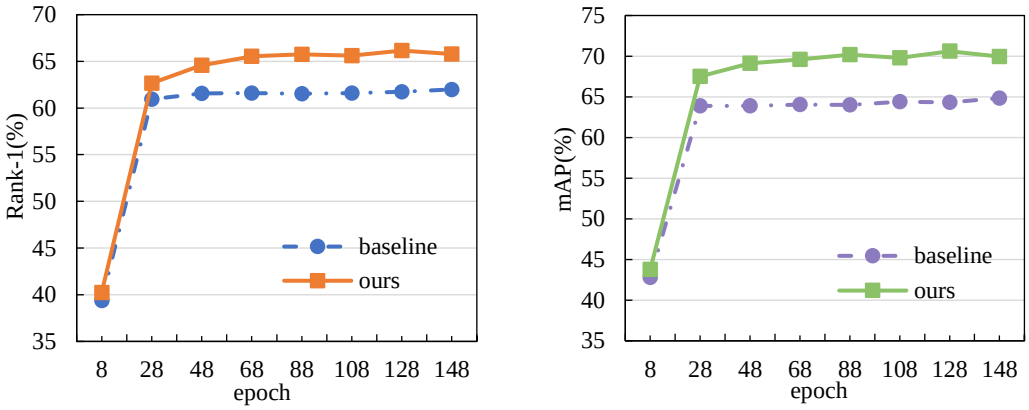
另外，在 SYSU-MM01 和 RegDB 数据集的全搜索和 Visible to Infrared 设置下，与 MCLNet[76]相比本章方法 R-1 分别高出 4.58%、5.08%，mAP 分别高出 4.73%、5.37%。虽该方法与本章方法均通过学习模态不变特征来缩小网络的模态差异，但不同的是，MCLNet 集中于混淆网络对模态的识别来实现在模态无关视角下的学习，忽视了网络训练数据不足的问题。本章方法利用 DEE 模块在骨干网络的共享层上生成更多的特征嵌入来丰富模态共享特征，增加了样本的训练数据，同时联合异质中心共享三元组损失学习优化网络模态共享能力，从而提升网络的精准度和可靠性。与 SFANet[55]相比本章方法 R-1 分别高出 4.24%、9.07%；mAP 分别高出 5.88%、10.44%；SFANet 提出利用灰度光谱图像完全取代 RGB 图像进行行人识别，能够在一定程度上减少模态差异，但由于 RGB 图像中颜色等细节信息丢失而导致网络精度下降。本章方法在不丢失颜色信息的基础上，从多样化共享特征信息融合的角度来捕获更多的特征信息，使网络具有更好的识别能力。次优的 PMT[77]方法利用灰度图像帮助网络学习，提出了一种渐进式学习策略，并通过多种损失函数的监督来增强网络所提取特征的鲁棒性。该方法与本章方法均以 AGW 为基础框架进行改进，与之相比，本章方法在 SYSU-MM01 数据集的全搜索模式下 R-1 和 mAP 分别高出 2.89%、1.46%，获得较优的识别准确度。说明本章方法利用 DEE 模块与损失函数联合优化基线网络，对模态不变特征的识别更为有效。

SFANet 方法和 PMT 方法均利用额外的灰度图像辅助网络进行特征学习，增加了网络的数据量和计算复杂度。本章方法仅对网络框架进行修改，利用多分支进行样本训练数据扩充，未引入额外训练数据，同时与基线相比在精准度提升的基础上网络训练时间不变，计算复杂度不高。

4.4.4 实验分析

本章方法与基线方法保持相同参数前提下，在 SYSU-MM01 数据集中对比了改进前后的首张图片命中率 Rank-1 和平均精度 mAP 的变化关系，由图 4-3 中

（a）和（b）所示，在训练阶段本章所提网络的 Rank-1 和平均精度大幅度高于基线网络，最终实验结果与基线网络相比获得有效地提升。



（a）　　（b）

图 4-3 训练阶段不同评价指标变化图。（a） Rank-1 变化过程。（b） mAP 变化过程。

表 4-3 和 4-4 展现了本章方法与基线方法在两个数据集不同模式下的实验数据，由图可以看出改进后的网络识别精度均高于基线方法。其中，本章算法在 SYSU-MM01 数据集的全搜索模式下 R-1、mAP、mINP 分别高出 5.23%、5.78%和 7.11%；在 RegDB 数据集的可见光到红外设置下 R-1、mAP、mINP 分别高出 2.99%、4.85%和 5.84%。实验结果表明所提网络模型中的 DEE 模块能够使骨干网络生成更多的特征嵌入，弥补原始网络模态共享信息的不足；而异中心共享三元组损失在同时处理两个模态的类内类间变化的基础上学习共享特征进一步缩小模态差异从而提升模型精度。

表 4-3 在 SYSU-MM01 数据集下的精准度对比实验结果（%）

method	All-search			Indoor-search		
	R-1	mAP	mINP	R-1	mAP	mINP
Baseline	64.75	60.93	45.72	70.65	75.39	71.19
ours	69.98	66.71	52.83	77.38	80.75	77.13

表 4-4 在 RegDB 数据集下的精准度对比实验结果（%）

method	Visible to Infrared			Infrared to Visible		
	R-1	mAP	mINP	R-1	mAP	mINP
Baseline	82.39	73.59	57.20	79.97	71.41	53.04
ours	85.38	78.44	63.04	87.08	79.79	65.76

为考察所提方法各个模块的有效性，在 SYSU-MM01 数据集下进行一系列实验，实验中的所有参数设置相同。如表 4-5 所示，Baseline 表示基线网络，即

在 AGW 原始网络的基础上使用随即擦除和通道增强操作。Ours1 表示使用异质中心共享三元组损失替换原始网络中的加权三元组损失，与 Base 相比 R-1 分别增加了 2.84%、2.91%，mAP 分别增加了 1.86%、2.26%，网络精度明显提升。说明该损失能够辅助网络提取模态共享特征，缩小类内距离，扩大类间差异。Ours2 表示在 Ours1 基础上添加多层嵌入模块，与 Ours1 相比 R-1 分别增加了 2.39%、3.82%，mAP 分别增加了 3.92%、3.1%，说明该模块通过多分支生成更多的特征嵌入有效丰富了网络捕获的特征，从而提高了模型的性能。

表 4-5 消融实验结果（%）

method	L_{wrt}	L_{hcg_tri}	DEE	All-search			Indoor-search		
				R-1	mAP	mINP	R-1	mAP	mINP
Baseline	✓	×	×	64.75	60.93	45.72	70.65	75.39	71.19
Ours1	×	✓	×	67.59	62.79	47.03	73.56	77.65	73.42
Ours2	×	✓	✓	**69.98**	**66.71**	**52.83**	**77.38**	**80.75**	**77.13**

此外，对于 DEE 模块，为评估其在骨干网络 ResNet50 内的不同共享层后嵌入对网络性能的影响，分别在骨干网络第二层（stage-2）、第三层（stage-3）和第四层（stage-4）后融入 DEE 来拓展嵌入数据。由图 4-4 可以看出在第四层后加入 DEE 效果最好，说明在较深层的共享骨干网络中模态间隙较小，使 DEE 生成能力更强，有效缩小了生成嵌入与原始嵌入之间的距离。

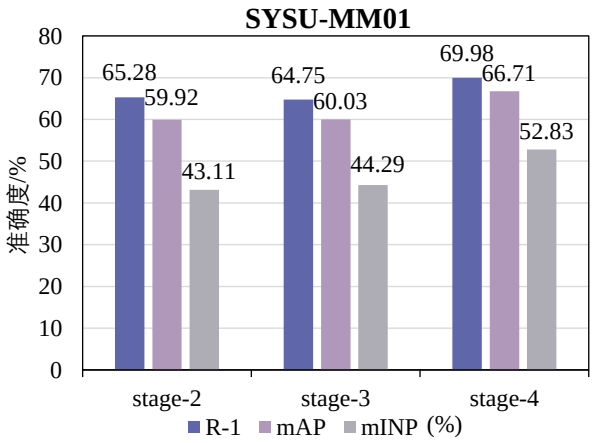


图 4-4 DEE 嵌入不同共享层的实验结果

为直观的展现算法的识别效果，从公开的大型数据集 SYSU-MM01 中分别挑选两位行人图片，采用最困难的全搜索模式 Single-hot 的设置下将基线和所提方法与目标行人检索匹配的前十个相似结果进行可视化，如图 4-5 所示。其中，query 表示待检索图像，行人图像上方的数字为绿色表示身份匹配正确，红色表

示身份匹配错误。从图中可以看出，Ours 的方法检索行人的准确率更高，侧面验证了改进算法的有效性。

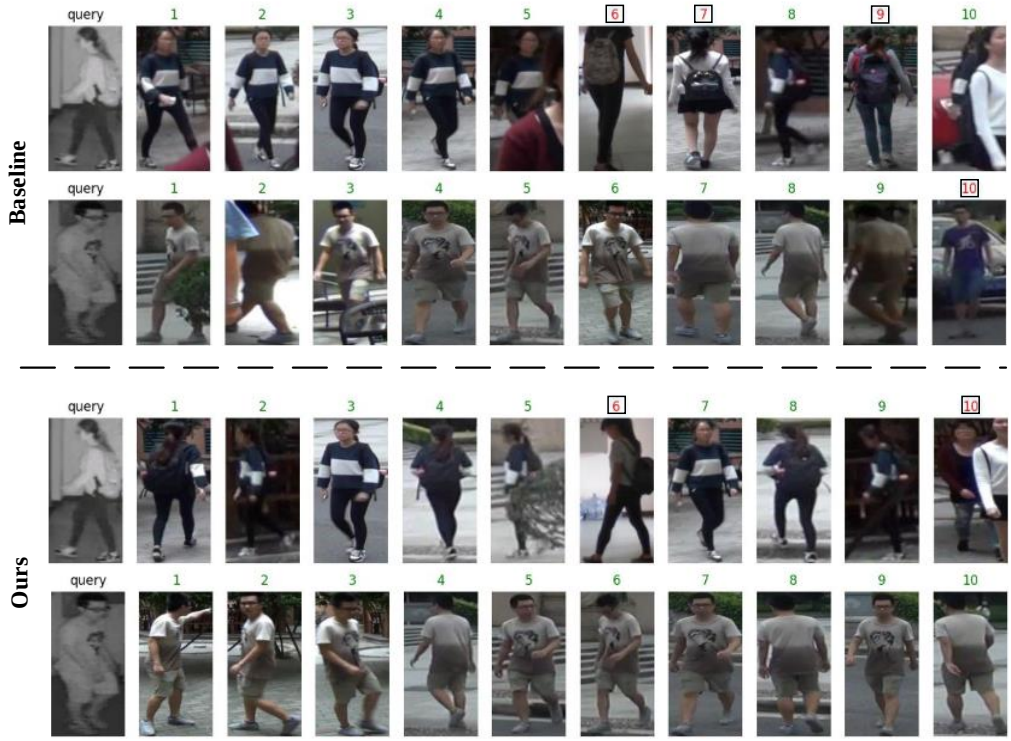


图 4-5 在 SYSU-MM01 数据集上的可视化结果图

4.5 本章小结

在本章中提出一种多样化深度信息融合嵌入的跨模态行人重识别改进方法。该方法在深层骨干网络构建多样化嵌入拓展模块，利用多分支生成并融合多种深度特征信息，拓展特征嵌入空间，实现了深层网络中模态共享信息的补充。同时利用异质中心共享三元组损失和交叉熵损失对扩充后的训练数据进行联合优化，进一步提升网络的表征能力及分类准确度。在两个公共数据集上进行了对比实验、消融实验和可视化实验，其中在 SYSU-MM01 数据集的全搜索模式下 R-1、mAP、mINP 分别高出 5.23%、5.78%和 7.11%；在 RegDB 数据集的可见光到红外设置下 R-1、mAP、mINP 分别高出 2.99%、4.85%和 5.84%。与现有的跨模态行人重识别算法比较，本章算法具有较高的准确度。实验结果表明，本章方法在一定程度上缓解了由网络缺乏有效训练数据导致提取表征能力不足的问题，在城市安全、社区安防等方面，有效提高了智能安防系统中模型的泛化能力和识别精度。

第 5 章 融合多尺度特征与混淆学习的跨模态行人重识别

5.1 引言

在复杂环境下，智能监控系统利用计算机视觉获取高级语义信息，以实现对特定行人身份的准确检索。在实际应用中，由于不同时间地点中行人图像数据受前景背景、光线视角以及姿势变化等影响，导致智能监控系统中跨模态 Re-ID 研究成为最大的挑战。研究存在的困难主要来自行人图像间的模态内和模态间差异。在大型商场、工业园区、商业街等人流密集的多人交互场景中，模态内差异主要由行人的姿态错位、障碍物遮挡、拍摄视角变化等因素引起。模态间差异是由相机原始的成像方式引起的模态差异，该差异使得可见光图像中如行人衣服颜色等重要的判别信息在红外图像中丢失，增加了识别的难度。在智能监控系统中，为了提高该类技术处理图像中行人姿态错位、障碍物遮挡以及颜色变化等干扰的适应能力。如何有效挖掘鉴别性信息应对模态内差异，以及提取不随模态而改变的信息应对跨模态变化以提升跨模态行人重识别的精度成为本章研究的关键。

论文第三、四章的算法均利用全局共享特征来缓解模态差异，由于实际监控场景常出现背景干扰，局部遮挡等情况，仅考虑全局信息忽略局部信息，容易导致关键细节信息的缺失，使网络的检索准确度下降。大多数方法通过对全局特征进行划分得到行人局部特征信息进行检索可以缓解上述难题，但当网络出现相似的不同行人时，上述方法忽略了从局部-全局信息互补的角度获取行人完整的特征表示，使其不足以分辨相似行人中不同的外观差异。此外，第三、四章算法专注于增强网络的表征能力，未采用有效的方法提取与模态不相关的信息，使网络难以应对图像间的跨模态变化。而常用的解决方法是利用生成对抗网络或图片风格转换的方式来缓解跨模态差异。虽然该方法在跨模态行人重识别网络中取得一定的效果，但常需要大量的成本，且生成图像不能完全取代对应的缺失图像，会带来额外的模态差异，从而对网络模型产生负面影响。

因此，针对上述问题，本章提出了一种基于多尺度特征与混淆学习的跨模态行人重识别改进方法，其目的是充分获取行人图像中具有判别性且与模式无

关的信息，以获得更强大的行人特征描述符。本章从全局特征与局部特征互补的角度出发，挖掘图像中有效的行人粗细粒度信息，获取丰富的鉴别性特征表示。同时，通过混淆模态识别反馈，将真伪模态标签信息与特征信息融合，利用对抗性学习的思想提取与模态无关的高级语义层特征来抑制模态差异，增强表示学习对模态变化的抗干扰能力。该策略直接作用于网络中，避免了额外噪声和模态差异的影响。最后，联合损失函数来监督网络进行更新和训练，进一步提高跨模态行人重识别模型的整体性能。

5.2 算法流程

网络结构如图 5-1 所示，主要由双流主干网络，多尺度特征互补模块，混淆学习策略（Confusion Learning Strategy, CLS）和损失函数几个部分组成。

双流网络采用 Resnet50 作为骨干来提取行人图像的特征，分别输入可见光图像和红外图像在第一层和第二层进行模态特定特征提取，这两层的网络参数独立。而骨干网络的后三层共享权重（share weights）作为特征嵌入部分，将网络提取的不同模态特征拼接后输入以提取模态共享特征，并且将最后一个卷积块的步幅由 2 改为 1，以获得较细粒度的行人特征。接着将网络分为双路径，分别聚焦于细腻和粗糙的特征区域，引入特征多样性。并采用混淆学习策略，确保优化专注于与模态无关的信息，从而提高共享特征表示的泛化能力。

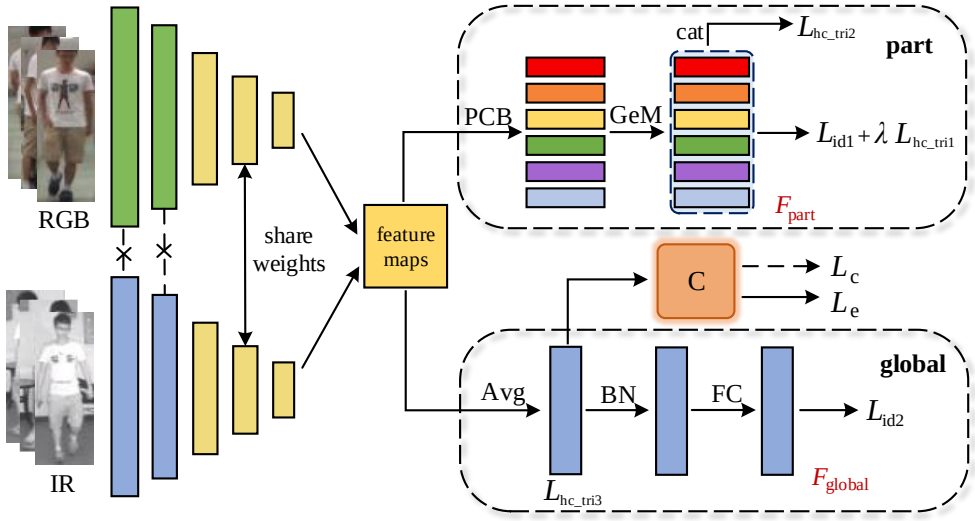


图 5-1 整体网络结构图

5.3 算法原理

5.3.1 多尺度特征互补模块

身体结构是一个人的固有特征，无论图像模态如何变化，如图5-2所示人的身体结构信息都是基于模态不变的，可以作为跨模态共享的判别信息来表示一个人。因此，为提取更多具有鉴别性的特征信息，以实现缩小模态差异的目的，本章设计了多尺度特征互补模块，采用局部深度特征和全局深度特征互补的形式来学习身体的局部结构与全局结构，挖掘有效的粗细粒度信息，提高特征的判别能力。

图 5-2 行人粗细粒度特征划分

如图5-1所示，网络由可见光和红外两条路径组成，分别提取来自两种模态的图像特征，图像在经过双流网络之后转化为三维特征图。分支结构设置如表5-1 所示。对于局部特征提取分支 part，采用 PCB（Part-based Convolutional Baseline）均匀划分策略。将三维特征图在水平方向均匀划分为 p 条，生成局部特征图以学习细粒度特征表示，引入内容粒度的多样性，其中 p=6。接着采用广义均值池化（Generalized-Mean, GeM）将三维的部分级特征转化为一维特征向量，并且使用 1×1 卷积块降低局部特征向量的维数，得到局部特征 F_{part} 。在分支的最后，将 p 块条纹特征连接（cat）用来描述人的身体结构，由异质中心三元组损失监督。对于全局深度特征提取部分 global，与部分级特征提取不同，该分支采用自适应平均池化（adaptive average pooling, Avg）来处理三维特征图，通过减小特征图大小来减少网络的计算量。然后，将特征向量输入批处理归一

化（Batch Normalization, BN）层进行归一化，并且采用维数为 2048 的全连接（Fully Connected, FC）层进行分类，将维度降至 512，可得到全局特征 F_{global}。最后，对于网络中提取的局部和全局特征采用异质中心三元组损失进行度量学习，然后采用标签平滑的识别损失对具有期望维数的全连接层输出的特征进行识别，而均匀划分的 p 块局部特征使用 p 个参数不共享的分类器。总的来说，多尺度特征互补模块使网络在捕获深层细化特征的同时也学习粗糙的全局深度特征，保持特征的多样性，以获取行人图像中完整的身体结构信息，从而增强行人的特征描述符。

表 5-1 局部和全局分支结构设置

Branch	Part	Dims	Feature
part	6	256	F_{part}
global	1	2048	F_{global}

5.3.2 混淆学习策略

由于可见光图像和红外图像来自两种不同的模态，难以进行特征对齐比较。为减少模态之间的差异，受 Xin 等人[76]的启发，将网络设置为忽略模态信息的结构，学习共享的特征表示。但共享的特征不等于有用的特征，故为了使网络在共有的视角下集中于收集有用的共享信息，而不过分关注琐碎的信息，本章采用一种混淆学习策略，利用对抗性学习的思路，混淆可见光和红外模态来欺骗网络，通过最小化不同模态间差异及最大化交叉模态相似度来实现特征对模态变化的鲁棒性。该方法直接嵌入在网络中，避免了生成图像质量差和多余噪声的影响，如图 5-3 所示。

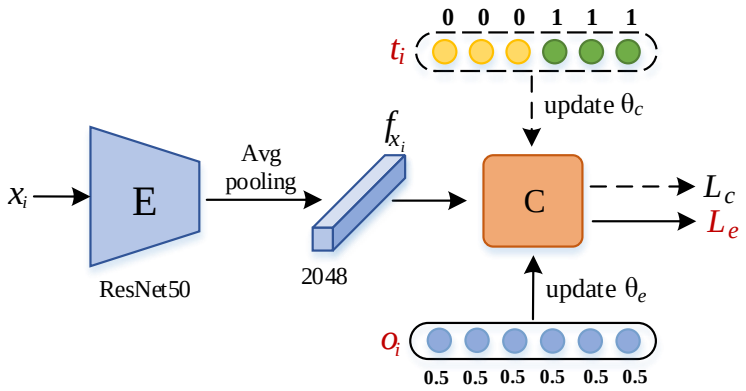


图 5-3 混淆学习结构图

具体地来说，混淆学习策略由特征提取器（E）和混淆模块（C）组成。每个样本图像 x_i 都有一个身份标签 y_i、一个真实的模态标签 t_i 和一个混淆的模态标签 O_i。对于每个输入的图像 x_i，使用一个二维向量来定义模态标签，将可见光图像和红外图像的真实模态标签分别设置为[1,0]和[0,1]，而来自两种模态所有样本混淆的模态标签 O_i 设置为[0.5,0.5]。混淆模块由两个分类器构成，对输入图像的模态进行分类，其使用参数 θ_c 来表示。将样本图像 x_i 经过特征提取器得到维度为 2048 的特征 f_{x_i} 输入混淆模块，并利用其得到的模态预测概率 $p_c(f_{x_i})$ 与真实的模态标签 t_i 进行比较，可得到混淆模块的损失函数表达式（5-1）。而特征提取器用来提取与模态无关且有鉴别性的特征。特征提取器为本章的 ResNet50 结构，用参数 θ_e 来表示。为了使特征提取器所捕获的特征达到混淆模态信息的作用，将特征提取器的预测概率与标签 O_i 进行比较，可得到公式（5-2）。

$$L_c(\theta_c) = -\frac{1}{N}\sum_{i=1}^{N} t_i \log p_c(f_{x_i}, \theta_c; \theta_e) \tag{5-1}$$

$$L_e(\theta_e) = -\frac{1}{N}\sum_{i=1}^{N} o_i \log p_c(f_{x_i}, \theta_e; \theta_c) \tag{5-2}$$

其中，N 表示一批中的样本数，x_i 表示第 i 个输入的样本图像，θ_e 和 θ_c 分别表示学习到的特征提取器和模态分类器，$p_c(f_{x_i}, \theta_c; \theta_e)$ 表示样本被正确分类的概率，$p_c(f_{x_i}, \theta_e; \theta_c)$ 表示提取到的样本特征能够实现模态混淆的概率，且公式（5-1）和（5-2）都使用 softmax 函数进行标准化。

在网络训练时，交替地更新 θ_e 和 θ_c 直到达到平衡，则可构成一个对抗损失，θ_e 和 θ_c 的优化如公式（5-3）、公式（5-4）和公式（5-5）所示。

$$L(\theta_c, \theta_e) = L_c(\theta_c) + L_e(\theta_e) \tag{5-3}$$

$$\hat{\theta}_c = \arg\min_{\theta_c} L(\theta_c, \hat{\theta}_e) \tag{5-4}$$

$$\hat{\theta}_e = \arg\min_{\theta_e} L(\hat{\theta}_c, \theta_e) \tag{5-5}$$

优化过程中，每次仅更新一个部分。即 θ_e 使行人的特征分布相似来增加混淆模块的损失。而 θ_c 通过不断地减少混淆模块分类器的损失，使网络能够正确辨别模态。最后，通过对抗训练实现特征提取器所捕获的特征使模态分类器不能够区分图像的模态。当网络实现混淆时，L_c 即可忽略。

5.3.3 损失函数

在本章提出的网络模型中，前部分通过多尺度特征学习和混淆学习操作得到了消除一定模态差异的特征。为了进一步处理跨模态和类内变化，保证行人图像分类的可行性和有效性，本章使用识别损失 L_{id} 和异质中心三元组损失 $L_{\mathrm{hc\_tri}}$ 联合监督，引导网络模型的训练。

识别损失的目的是对行人身份进行分类，通过将不同模态同一个人的特征视为同一类来整合行人的特征，学习模态特定特征来处理模态内变化。本章中采用标签平稳操作的识别损失，以防止模型训练过拟合，计算方式如（5-6）所示。

$$L_{\mathrm{id}} = \sum_{i=1}^{N} -q_i \log(p_i) \tag{5-6}$$

$$q_i = \begin{cases} 1 - \dfrac{N-1}{N}\xi, y = i, \\ \dfrac{\xi}{N}, y \neq i, \end{cases} \tag{5-7}$$

其中，N 为总训练集的身份个数。对于一个图像，y 表示为真实 ID 的标签，p_i 表示第 i^{th} 类的 ID 预测对数，ξ 是一个常数，用来减少模型对训练集的置信度。

在跨模态行人重识别任务中通常采用三元组损失作为度量学习，但由于三元组损失通过锚与其他样本的比较来计算损失，具有强大的约束力，可能存在一些异常值，对过于苛刻的限制成对距离造成不利影响。因此，本章采用异质中心三元组损失监督网络学习。该损失将三元组损失与中心损失的优势相结合，考虑使用每个人的中心作为身份，通过替换锚点中心与其他样本的比较来放松约束，从而缓解不利的影响。故，每个模态的身份特征中心的计算公式如（5-8）和（5-9）所示。

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i \tag{5-8}$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{5-9}$$

其中，v_j^i 表示第 i 个行人的第 j 张 RGB 图像，t_j^i 表示第 i 个行人的第 j 张 IR 图像，c_v^i 表示第 i 个行人 RGB 图像的中心，c_t^i 表示第 i 个行人 IR 图像的中心。

异质中心三元组损失不仅关注类内交叉模态紧凑型，还集中于挖掘类间可

分离性特征，弥补了识别损失不能够应对跨模态变化的缺点。基于 PK 采样的方法，在每个批次中都有 P 个 RGB 图像的中心 $\left\{c_v^i \middle| i=1,2,3\cdots,P\right\}$ 和 P 个 IR 图像的中心 $\left\{c_t^i \middle| i=1,2,3\cdots,P\right\}$。因此，在跨模态行人重识别中，基于 PK 采样策略和中心距离公式，异质中心三元组损失可表述为公式（5-10）。

$$L_{\text{hc\_tri}}=\sum_{i=1}^{P}\left[\rho+\left\|c_v^i-c_t^i\right\|_2-\min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_v^i-c_n^j\right\|_2\right]_+ +\sum_{i=1}^{P}\left[\rho+\left\|c_t^i-c_v^i\right\|_2-\min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|c_t^i-c_n^j\right\|_2\right]_+ \quad (5\text{-}10)$$

其中，c_n^j 为第 j 个行人的难负样本中心，ρ 为异质中心三元组损失的边缘值。每个身份，$L_{\text{hc\_tri}}$ 只关注一个跨模态的正对和在模态内、模态间挖掘最难的负对。

总的来说，对于局部特征分支，本章采用标签平滑的识别损失和异质中心三元组损失对细化特征进行处理，而拼接后的细化特征仅使用异质中心三元组损失监督，如公式（5-11）所示。对于全局特征分支，联合标签平滑的识别损失、异质中心三元组损失和混淆学习策略的对抗损失进行学习，如公式（5-12）所示。因此，本章的总体损失函数为公式（5-13）所示。

$$L_{\text{part}}=\sum_{i=1}^{p}\left(L_{\text{id1}}^i+\lambda L_{\text{hc\_tri1}}^i\right)+L_{\text{hc\_tri2}} \quad (5\text{-}11)$$

$$L_{\text{global}}=L_e+L_{\text{hc\_tri3}}+L_{\text{id2}} \quad (5\text{-}12)$$

$$L_{\text{all}}=L_{\text{part}}+L_{\text{global}} \quad (5\text{-}13)$$

其中，p 为部分级条纹特征数，λ 为预定义的权衡参数。

5.4 实验结果与分析

5.4.1 实验设置

本章实验环境：CPU 为 4 核 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存 16G，显卡 NVIDIA GeForce RTX2080Ti（显存 11G），操作系统为 Ubuntu 16.04，深度学习框架 pytorch1.1.0。行人图像为 288×144 大小，数据增强通过对图像进行随机裁剪和翻转实现。异质中心三元组损失预定义的边缘值设置为 0.3。采样策略中，SYSU-MM01 数据集下设置了 P=6，K=8，RegDB 数据集下设置了 P=8，K=4，即在 1 个批次中，随机选取 P 个行人身份，每个身份包含 K 张 RGB 图像和 K 张 IR 图像。对于参数 λ 的取值采取文献[65]中的设置，即在 SYSU-MM01 数据集中，λ=0.1，在 RegDB 数据集中，λ=0.2。部分级条纹特征 p 设置为

6。训练阶段 epoch 大小为 80；初始学习率为 0.1，在 20、50 个 epoch 时学习率衰减 0.1 和 0.01，前 10 个 epoch 应用热身策略；SGD 优化器动量参数为 0.9。

5.4.2 实验数据集与评价指标

本章实验采用 SYSU-MM01、RegDB 数据集进行验证，在测评时，采用 Rank-*n*、mAP、mINP 作为评价指标。

5.4.3 与其他方法进行比较

为验证本章算法的先进性，将本章所提算法与现有的跨模态行人重识别算法进行比较。在两个公共数据集 SYSU-MM01 和 RegDB 的实验结果如表 5-2 和表 5-3 所示，主要包括 BDTR[41]、Align GAN[51]、Xmodel[53]、TSLFN[43]、HCT[65]、MMN[54]、MAUM[50]等算法。其中“—”表示原论文中没有报告的结果。

表 5-2 在 SYSU-MM01 数据集和其他方法对比实验结果（%）

method	All-search					Indoor-search				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[39]	14.80	54.12	71.33	15.95	—	20.58	68.38	85.79	26.92	—
TONE[40]	12.52	50.72	68.60	14.42	—	20.82	68.86	84.46	26.38	—
HCML[40]	14.32	53.16	69.17	16.16	—	24.52	73.25	86.73	30.08	—
BDTR[41]	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
eBDTR[48]	27.82	67.34	81.34	28.42	—	—	—	—	—	—
D²RL[66]	28.90	70.60	82.40	29.20	—	—	—	—	—	—
MAC[67]	33.26	79.04	90.09	36.22	—	33.37	82.49	93.69	44.95	—
DPMBN[69]	37.02	79.46	89.87	40.28	—	44.17	87.12	95.24	54.51	—
Align GAN[51]	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
LZM[71]	45.00	89.06	95.77	45.94	—	54.28	94.22	98.22	63.92	—
Xmodal[53]	49.92	89.79	95.96	50.73	—	—	—	—	—	—
DDAG[44]	54.75	90.39	95.81	53.02	—	61.02	94.06	98.41	67.98	—
TSLFN[43]	59.96	91.50	96.82	54.95	—	59.74	92.07	96.22	64.91	—
HCT[65]	61.68	93.10	97.17	57.51	39.54	63.41	91.69	95.28	68.17	64.26
cm-SSFT[75]	61.60	89.20	93.90	63.20	—	70.50	94.90	97.70	72.60	—

TSME[78]	64.23	95.19	98.73	61.21	—	64.80	96.92	99.31	71.53	—
MMN[54]	70.60	96.20	99.00	66.90	—	76.20	97.20	99.30	79.60	—
MPANet[79]	70.58	96.21	98.80	68.24	—	76.74	98.21	99.57	80.95	—
MAUM[50]	71.68	—	—	68.79	—	76.97	—	—	81.94	—
RAPV[80]	63.97	95.30	98.70	62.33	49.34	69.00	97.39	99.38	75.41	71.94
第三章算法	56.36	90.29	95.69	54.96	42.01	62.20	94.26	97.74	69.11	65.50
第四章算法	69.98	96.29	98.80	66.71	52.83	77.38	96.85	99.17	80.75	77.13
Ours	**76.69**	**94.89**	**97.48**	**72.45**	**59.64**	**81.19**	**94.32**	**97.27**	**82.78**	**79.42**

表 5-3 在 RegDB 数据集和其他方法对比实验结果（%）

method	**Visible to Infrared**					**Infrared to Visible**				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[39]	17.75	34.21	44.35	18.90	—	16.63	34.68	44.25	17.82	—
HCML[40]	24.44	47.53	56.78	20.08	—	21.70	45.02	55.58	22.24	—
BDTR[41]	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
eBDTR[48]	34.62	58.96	68.72	33.46		34.21	58.74	68.64	32.49	
MAC[67]	36.43	62.36	71.63	37.03	—	36.20	61.68	70.99	36.63	—
Align GAN[51]	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[53]	62.21	83.13	91.72	60.18	—	—	—	—	—	—
HCT[65]	91.05	97.16	98.57	83.28	68.84	89.30	96.41	98.16	81.46	64.81
cm-SSFT[75]	72.30	—	—	72.90	—	71.00	—	—	71.70	—
MMN[54]	91.60	97.70	98.90	84.10	—	87.50	96.00	98.10	80.50	—
MPANet[79]	83.70	—	—	80.90	—	82.80	—	—	80.70	—
MAUM[50]	87.87	—	—	85.09	—	86.95	—	—	84.34	—
RAPV[80]	86.81	95.98	97.86	81.02	68.19	86.60	96.14	97.86	80.52	65.73
第三章算法	78.88	91.36	94.88	70.74	55.37	76.74	89.77	93.13	68.71	51.78
第四章算法	85.38	95.26	97.42	78.44	63.04	87.08	96.27	98.03	79.79	65.76
Ours	**94.62**	**97.20**	**98.72**	**94.60**	**92.96**	**92.77**	**96.17**	**98.16**	**93.39**	**91.95**

从表 5-2、表 5-3 可以看出，本章算法在两个公共数据集上取得了较高的精度，均优于对比方法。具体来看，TONE[40]、BDTR[41]、eBDTR[48]等以往的方法主要利用双流网络来提取红外图像和可见光图像的模态特定特征，并通过改进损失函数来约束行人特征的距离，但网络忽略了模态共享特征对处理模态间变化的重要性。本章方法将网络的主干部分划分为参数特定和参数共享的两个阶段，以分别提取行人的模态特异性信息和模态共享信息，为深层的网络学习

特征表示奠定了基础。D^{2}RL[66]、Align GAN[51]、Xmodal[53]、MMN[54]等方法利用转换图片风格的方式来减轻模态差异带来的影响，但生成的图像不仅扩大了工作量还容易引入新的噪声干扰。而本章算法在不混入多余噪声的前提下利用最大-最小博弈的方法，混淆网络提取两种模态的特征，迫使网络专注于优化与模态无关的共享特征，同样缓解了模态差异的影响。由实验对比结果可知，与此类方法中效果最好的 MMN 算法相比，所提方法在 SYSU-MM01 数据集全搜索设置下 R-1 结果提升 6.09%，mAP 结果提升 5.55%；在 RegDB 数据集的可见光到红外设置下 R-1 结果提升 3.02%，mAP 结果提升 10.5%。TSLFN[43]、HCT[65]等方法通过对行人特征均匀划分进行细化学习，来最大化跨模态特征的相似性。如 HCT 算法对提取的特征进行切块处理，使网络仅关注行人图片中的细粒度信息，并在此基础上联合基于样本中心的损失函数来监督学习，通过拉近同一个人的特征距离，推远不同人的特征距离来提高行人识别的准确率。但由于网络过于依赖细节信息，而忽略了对整体特征的把握，导致模型精准度不高。本章网络将细粒度局部特征与粗粒度全局特征相结合，增加特征的丰富度来弥补这一缺陷，使模型精度得到明显提升。在 SYSU-MM01 数据集与 HCT 方法相比，全搜索设置下本章算法的 R-1、mAP 和 mINP 分别高出 15.01%、14.94%和 20.10%；在 RegDB 数据集与 HCT 方法相比，可见光到红外设置下本章算法的 R-1、mAP 和 mINP 分别高出 3.57%、11.32%和 24.12%。与次优的 MAUM[50]方法相比，在 SYSU-MM01 数据集的全搜索设置下和 RegDB 数据集可见光到红外设置下，Rank-1 分别增长了 5.01%、6.75%，mAP 分别提升了 3.66%、9.51%，充分验证了本章方法的先进性。与 Xmodal、MMN 等利用转换图片风格的方法相比，本章方法利用混淆学习策略缩小模态差异，该方法直接嵌入在网络中，未引入额外训练数据，计算复杂度不高，同时避免了上述方法生成图像质量差和多余噪声的影响。对于次优的 MAUM 方法，虽然获得了一定的优越性，但从实际应用的角度来看，与本章方法相比工作量较大，且不适用于大规模数据集。

此外，将本章算法在公共的数据集与第三章和第四章算法进行比较。基于相同的评估指标，对比第三、四章算法，在 SYSU-MM01 数据集全搜索模式下本章算法 mAP 分别提升 17.49%、5.74%，在 RegDB 数据集可见光到红外设置下本章算法 mAP 分别高出 23.86%、16.16%。由此分别看来，第三章深度信息融

合变分细化蒸馏的跨模态行人重识别方法采用双分支学习并融合网络中不同粒度的全局特征信息，再对学习到的数据信息进行去噪处理，虽然在一定程度上提升了基线算法的性能，但在网络训练数据样本有限的情况下，该算法使用双分支提取更多特征的同时采用变分细化蒸馏策略压缩数据进行去除噪声，可能对次显著的关键信息进行削弱，导致模型学习能力不足，局限了该算法的性能。第四章基于多样化深度信息融合嵌入的跨模态行人重识别方法在第三章基线算法的基础上引入了数据增强策略扩充训练数据，联合了特征嵌入和度量学习，依附多样化嵌入拓展模块的多分支结构融合新生成与原始网络的特征信息丰富训练数据以获取充分的特征表示；同时借助异质中心共享三元组损失约束行人的类内类间距离，进一步提升网络性能。从整体上来看，第三、四章算法均在全局特征的基础上进行数据信息的学习和融合，更关注行人主干部位的信息提取往往忽略了四肢和边缘的局部信息。本章算法采用多尺度的特征互补网络以分别学习行人图像中的粗粒度全局和细粒度局部信息保证了行人信息的完整度；混淆学习策略将真伪标签信息与特征信息融合，利用对抗性学习的思路混淆了网络的模态识别反馈，学习无关于模态的高级语义信息，提升了网络应对图像跨模态变化的能力。二者联合辅助模型增强对不同身份行人的相似细节特征的精准识别，从而提升模型识别精度。

5.4.4 实验分析

表 5-4 和表 5-5 展现了本章方法与基线方法在 SYSU-MM01 和 RegDB 数据集中的性能对比数据，可以看出所提方法较基线方法在 SYSU-MM01 数据集全搜索模式和 RegDB 数据集可见光到红外模式下 Rank-1 分别提升了 5.37%、1.21%，mAP 分别提升了 5.86%，0.95%，mINP 分别提升了 7.96%、0.98%。实验结果表明采用多尺度特征的互补学习有助于形成完整的行人表征，混淆学习有利于缩小行人特征间的跨模态差异，从而帮助网络性能提高。

表 5-4 在 SYSU-MM01 数据集下的精准度对比实验结果（%）

method	All-search			Indoor-search		
	R-1	mAP	mINP	R-1	mAP	mINP
Baseline	71.32	66.59	51.68	76.75	78.46	74.37
ours	76.69	72.45	59.64	81.19	82.78	79.42

表 5-5 在 RegDB 数据集下的精准度对比实验结果（%）

method	Visible to Infrared			Infrared to Visible		
	R-1	mAP	mINP	R-1	mAP	mINP
Baseline	93.41	93.65	91.98	91.74	92.62	91.19
ours	94.62	94.60	92.96	92.77	93.39	91.95

为评估网络结构中各个模块对网络的影响，在 SYSU-MM01 数据集的两种模式下进行了一系列消融实验，结果如表 5-6 所示。具体来说 Baseline 表示基线，即在 HCT 网络的基础上采用 k-reciprocal[81]重排序操作；global 表示全局特征分支；CLS 表示在网络中使用混淆学习策略。

由表 5-6 可知，在 Base 基础上将全局特征分支与部分级特征分支共同作用，进一步提升了网络模型的性能。即在全搜索模式下 R-1 提升了 2.62%，mAP 提升了 4.76%，mINP 提升了 4.47%；在室内搜索模式下，R-1 提升了 2.93%，mAP 提升了 2.98%，mINP 提升了 4.23%，可见该方法提取到不同粒度特征能够使网络学习到更多具有鉴别性信息的特征，从而获得更具鲁棒性的行人表示，有效地对行人图像进行识别分类。接着在全局特征提取部分融入混淆学习策略，使网络在全搜索模式下 R-1、mAP 和 mINP 分别提高了 2.75%、1.1%和 3.49%；在室内搜索模式下检索性能也得到进一步提高。说明网络通过最大最小博弈的方法，能够使优化集中于与模式无关且有用的特征上，增强了表示学习对模态变化的泛化能力。总的来说与基线相比，本章方法在 SYSU-MM01 数据集的全搜索模式下 R-1、mAP 和 mINP 结果分别提升了 5.37%、5.86%和 7.96%；在室内搜索模式下 R-1、mAP 和 mINP 分别提升了 4.44%、4.32%和 5.05%。

表 5-6 SYSU-MM01 数据集下的消融实验（%）

method	All-search			Indoor-search		
	R-1	mAP	mINP	R-1	mAP	mINP
Baseline	71.32	66.59	51.68	76.75	78.46	74.37
Baseline+global	73.94	71.35	56.15	79.68	81.44	78.60
Baseline+global+CLS	76.69	72.45	59.64	81.19	82.78	79.42

此外，为进一步探索全局分支中不同池化方式和不同特征进行混淆学习对网络性能的影响，在 SYSU-MM01 数据集下进行了一系列实验。

如图 5-4 所示为使用广义均值池化（GeM）、最大池化（Max）和自适应平均池化（Avg）来分别构建全局分支。由实验结果可以看出，使用自适应平均

池化的网络平均精度与平均逆置负样本惩罚率最高，说明平均池化能够代替其它池化来整合全局空间信息，增强对输入空间变化的鲁棒性，从而提升网络模型的精度，故所提方法采用自适应平均池化构建全局分支。

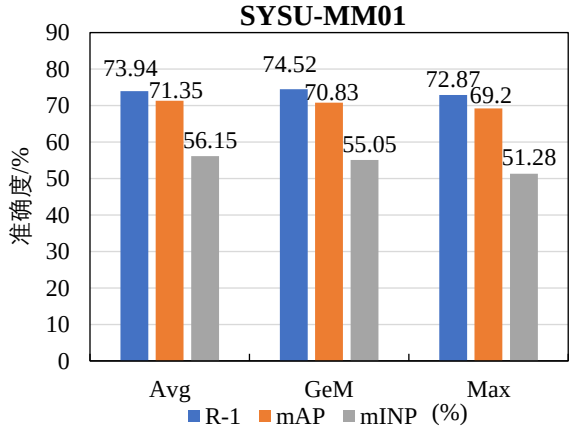


图 5-4 网络中采用不同池化方式的性能实验结果

表 5-7 为分别对池化（Pooling）层、批量归一化（BN）层和全连接（FC）层得到的行人特征进行混淆学习。具体的，将各层输出的特征作为混淆学习策略中特征提取器的输出来测试网络的性能。如表所示，在全搜索模式和室内搜索模式下的各个指标都表明对池化（Pooling）层输出的特征进行混淆学习效果最佳，而 BN 层和 FC 层输出的特征均难以实现模态混淆学习的目的，侧面验证了本章选择在 ResNet50 后靠前的高级语义层进行模态混淆学习的有效性。

表 5-7 模态混淆模块不同位置对模型性能影响的实验结果（%）

method	all-search			Indoor-search		
	R-1	mAP	mINP	R-1	mAP	mINP
BN	73.25	70.49	58.35	79.21	81.97	78.63
FC	75.29	71.06	57.99	78.23	80.55	77.08
Pooling	**76.69**	**72.45**	**59.64**	**81.19**	**82.78**	**79.42**

为验证检索结果的改进情况，从公开的大型数据集 SYSU-MM01 中分别挑选两位行人图片，采用最困难的全搜索模式 Single-hot 的设置下将基线和所提方法的前十个与目标行人相似的排序结果可视化，如图 5-5 所示。其中，query 表示待检索图像，图像上方从 1 到 10 的绿色数字代表身份匹配正确，框出来的红色数字表示身份匹配错误。从检索结果可以看出本章改进算法能够有效识别行人身份，检索出正确的结果，提高了待检索图像的匹配度。

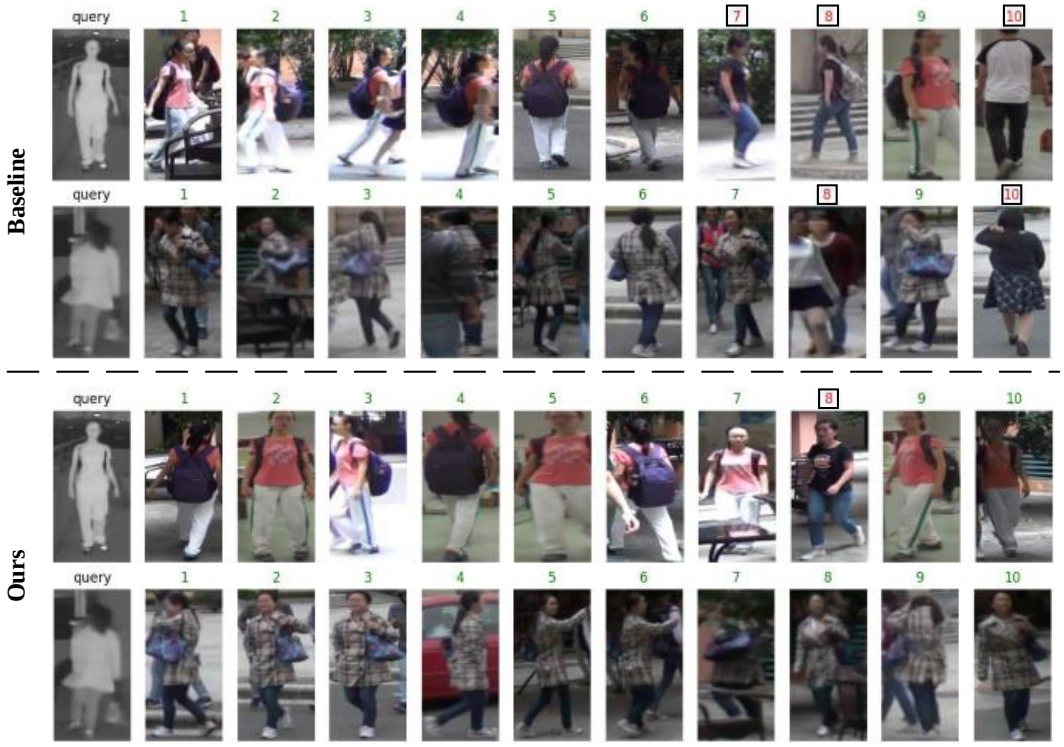


图 5-5 在 SYSU-MM01 数据集上的可视化结果图

5.5 本章小结

本章从充分利用行人的跨模态特征和缓解模态差异的角度出发，针对跨模态行人重识别任务提出了一种基于多尺度特征与混淆学习的改进网络，兼顾学习具有鉴别性和与模态无关的特征。具体地，网络通过多尺度特征互补模块同时捕获不同尺度的语义信息，丰富了行人特征图的可鉴别性，提高了特征表示的泛化能力。同时考虑到模态变化，利用混淆学习将模态标签信息与特征信息融合以混淆不同模态的特征输入，使网络在与模态无关的视角下学习有用的特征，提高了表征对跨模态差异的抗干扰能力。最后，在损失函数的引导下，完成了网络的进一步优化。在大规模数据集 SYSU-MM01 全搜索模式下 Rank-1 提升了 5.37%，平均精度提升了 5.86%。结果表明，该方法通过学习不同尺度的语义信息提高了模型的表征学习能力，使其能更准确地识别不同身份行人之间相似的细节特征，提高算法的泛化能力。在部署大型的智能监控系统时，本章算法可以很好地适应人流量密集的场景，减少因行人姿态错位、遮挡、和背景杂波等因素引起的行人搜索匹配度不高的问题，同时模型专注于优化模态共享的特征，缓解了跨模态图像颜色变化对其性能的影响。

第 6 章 总结与展望

6.1 工作总结

视频监控遍布生活各处构成宏大的监控网络，其中视频信息推动了公共安防、视频分析等领域的发展。行人重识别作为视频分析的常用技术，在可见光条件下的研究已取得卓越的成效。但实际公共场合的监控已不局限于使用可见光摄像机进行视频监控追踪。为保证智能监控的全天候运作，夜间可切换红外模式的监控相机被广泛应用，提出针对白天可见光夜间红外的跨模态行人重识别任务。该任务旨在研究在不同光照环境、不同场景或地点下进行的行人检索匹配。而如何减少可见光和红外图像间的跨模态差异，如何获取具有鉴别性、鲁棒性的行人表征缩小模态内差异从而实现更精准的识别为该技术研究重点。因此，本文从这两个角度出发，基于深度信息融合、特征学习和度量学习等方面展开研究，具体的研究内容总结如下：

（1）论述了单模态和跨模态行人重识别技术的背景、意义以及存在的难题，对其在现实场景中实现的技术流程、国内外不同研究方法和相关基础知识进行了分析和总结。

（2）从深度信息融合和特征学习的角度出发，提出了深度信息融合变分细化蒸馏的跨模态行人重识别改进方法，采用双重信息聚合模块和变分细化蒸馏策略相结合。设计了多分支结构，分别采用广义均值池化增强对细化特征的关注度，再将多分支中不同的特征信息融合，使网络共同学习图像中细腻和粗糙的特征区域，提高了行人的表征判别能力；此外，为强化行人特征的鲁棒性，采用变分细化蒸馏策略，对多分支提取的特征信息进行精炼学习，在最大化有用信息提取的同时弱化了噪声等冗余的信息对网络的影响，提高了模型的稳定性。解决了智能监控中由多视角变化以及多模态特征变化引起的挖掘重要特征信息不足和识别率降低的问题。

（3）从多样化深度特征信息融合和度量学习的角度出发，提出了多样化深度信息融合嵌入的跨模态行人重识别改进方法。为缓解现有网络模型中有效训练数据匮乏的问题，在骨干网络的共享特征提取阶段，采用多样化嵌入拓展模

块学习多样化深度信息，通过多分支膨胀卷积扩大感受野拼凑为新的特征图再与上一阶段输入的特征信息融合来拓展特征空间，丰富了网络训练数据的多样性；为学习具有更强泛化能力的特征，优化了网络的度量学习策略，利用交叉熵损失和异质中心共享三元组损失进行训练学习，进一步增强了对类内、类间特征距离的约束，辅助网络提升了对交叉模态特征的学习能力。该方法有效解决了由缺少数据样本引起的模型学习能力不足问题，提高了智能安防系统中模型的分类能力和识别精度。

（4）从特征学习和标签与特征信息融合的角度出发，提出了融合多尺度特征与混淆学习的跨模态行人重识别改进方法。采用多尺度特征互补模块和混淆学习策略联合，兼顾学习具有鉴别性和模态无关的特征。其中多尺度特征互补模块学习有效的行人不同粒度信息，保持了特征的多样性和行人信息的完整度；此外，在不混入多余噪声的前提下，设计了将模态标签信息与特征信息融合的混淆学习策略，通过对抗性学习训练，实现特征提取器所捕获的特征使模态分类器不能够区分图像的模态，以达到混淆，强化了网络在模态无关的视角下对有用特征的关注度。在智能监控系统中，该算法可以对不同身份行人的相似细节特征进行更精准识别，同时增强了模型应对行人跨模态特征差异的鲁棒性，提高了其在现实智能监控中应对遮挡、行人错位及图像颜色变化的能力。

6.2 工作展望

跨模态行人重识别作为行人重识别领域的重要研究方向目前仍处于发展阶段。在复杂的现实街头和安防领域严苛的识别精准度要求下，跨模态行人重识别将面临很多挑战。虽然论文所提方法在一定程度上改进了跨模态行人重识别模型性能，但仍有很多不足之处，有待深入探究与突破。

（1）构建规模更大的跨模态公共数据集。论文算法采用的数据集是 SYSU-MM01 和 RegDB，由于数据集规模不大，存在一定局限性。在跨模态行人重识别研究领域，公共数据集较少，训练样本的不充足也限制了跨模态行人重识别的发展。因此，迫切需要新的大规模高质量的数据集，在多个跨模态数据集进行训练提高模型的泛化能力，缩小与现实监控系统数据量之间的差异。

（2）基于无监督的跨模态行人重识别研究。论文研究基于有监督的行人重

识别方法，所使用数据集的训练数据中均包含标签信息。而在实际监控环境下，检索的行人图像没有数据标注，考虑到跨模态行人重识别的实用性和应用前景，基于无监督的跨模态行人重识别研究更贴近实际应用。

（3）融入姿态预测网络。论文研究是基于智能监控系统的行人重识别，对行人进行定位追踪、抓捕犯罪嫌疑人，维护社会安全。未来可以在此基础上添加姿态预测网络融入行人的姿态特征来丰富模型提取的关键特征信息，同时结合姿态预测网络提前对行为可疑人员进行追踪定位，预防违法犯罪等事件发生。

参考文献

[1] 李擎, 胡伟阳, 李江昀, 等. 基于深度学习的行人重识别方法综述[J]. 工程科学学报, 2022, 44(5): 920-932.

[2] 王素玉, 肖塞. 行人重识别研究综述[J]. 北京工业大学学报, 2022, 48(10): 1100-1112.

[3] An F P, Wang J R, Liu R J. Pedestrian re-identification algorithm based on attention pooling saliency region detection and matching[J]. IEEE Transactions on Computational Social Systems, 2024, 11(1): 1149-1157.

[4] Huang B L, Piao Y, Tang Y F. An adaptive partitioning and multi-granularity network for video-based person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5282-5295.

[5] Chai T R, Chen Z Y, Li A N, et al. Video person re-identification using attribute-enhanced features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7951-7966.

[6] Asperti A, Fiorilla S, Orsini L. A generative approach to person reidentification[J]. Sensors, 2024, 24(4): 1240.

[7] Gao H, Hu C C, Han G, et al. Point-level feature learning based on vision transformer for occluded person re-identification[J]. Image and Vision Computing, 2024, 143: 104929.

[8] Qi B G, Chen Y, Liu Q, et al. An image-text dual-channel union network for person re-identification[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-16.

[9] Wang Z J, Xue J Y, Wan X L, et al. ASPD-Net: Self-aligned part mask for improving text-based person re-identification with adversarial representation learning[J]. Engineering Applications of Artificial Intelligence, 2022, 116: 105419.

[10] Krizhevsky A, Sutskever I, et al. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[11] Zhang W B, Li Z Y, Du H S, et al. Dual-stream feature fusion network for person

re-identification[J]. Engineering Applications of Artificial Intelligence, 2024, 131: 107888.

[12] Li J, Zhang S, Tian Q, et al. Pose-guided representation learning for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 622-635.

[13] 张红颖, 王徐泳, 彭晓雯. 结合前景分割的多特征融合行人重识别[J]. 中国图象图形学报, 2023, 28(05): 1360-1371.

[14] Yu Z, Qin W, Huang Z, et al. Joining features by global guidance with bi-relevance trihard loss for person re-identification[J]. Neural Computing and Applications, 2022, 34(11):8697-8712.

[15] Huang Z, Guan T, Qin W, et al. Gaussian-based probability fusion for person re-identification with Taylor angular margin loss[J]. Neural Computing and Applications, 2022, 34(23): 20639-20653.

[16] Gu H, Li J, Fu G, et al. AutoLoss-GMS: Searching generalized margin-based softmax loss function for person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 4744-4753.

[17] Mignon A, Jurie F. Pcca: A new approach for distance learning from sparse pairwise constraints[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012: 2666-2672.

[18] Zheng W S, Gong S, Xiang T. Reidentification by relative distance comparison[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(3): 653-668.

[19] Liu C, Loy C C, Gong S, et al. POP: Person re-identification post-rank optimisation[C]. Proceedings of the IEEE International Conference on Computer Vision, Portland, Oregon, US, 2013: 441-448.

[20] Li Z, Chang S, Liang F, et al. Learning locally-adaptive decision functions for person verification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, US, 2013: 3610-3617.

[21] Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C]. Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 1367-1376.

[22] Wang F Q, Zuo W M, Lin L, et al. Joint learning of single-image and cross-image representations for person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 1288-1296.

[23] Si J, Zhang H, Li C G, et al. Dual attention matching network for context-aware feature sequence based person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 5363-5372.

[24] Sun Y Y, Zheng L, Li Y L, et al. Learning part-based convolutional features for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 902-917.

[25] Sun Y, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 480-496.

[26] Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]. Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 2018: 274-282.

[27] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 1062-1071.

[28] 邵晓雯, 帅惠, 刘青山. 融合属性特征的行人重识别方法[J]. 自动化学报, 2022, 48(02): 564-571.

[29] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019, 95(11): 151-161.

[30] 朱敏, 明章强, 闫建荣, 等. 基于生成对抗网络的行人重识别方法研究综述[J]. 计算机辅助设计与图形学学报, 2022, 34(02): 163-179.

[31] Matsukawa T, Suzuki E. Person re-identification using CNN features learned from combination of attributes[C]. Proceedings of the 23rd International Conference on

Pattern Recognition, Cancun, 2016: 2428-2433.

[32] Su C, Zhang S, Xing J, et al. Deep attributes driven multi-camera person re-identification[C]. Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 475-491.

[33] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 3754-3762.

[34] Ding S Y, Lin L, Wang G R, et al. Deep feature learning with relative distance comparison for person re-identification[J]. Pattern Recognition: The Journal of the Pattern Recognition Society, 2015, 48(10): 2993-3003.

[35] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[36] Chen W H, Cheng X T, Zhang J G, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 1320-1329.

[37] Zhao C R, Lv X B, Zhang Z, et al. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(12): 3180-3195.

[38] Lou H, Jiang W, Gu Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2597-2609.

[39] Wu A C, Zheng W S, Yu H X, et al. RGB-infrared cross-modality person re-identification[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 5390-5399.

[40] Ye M, Lan X Y, Li J W, et al. Hierarchical discriminative learning for visible thermal person re-identification[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, Lousiana, USA, 2018: 7501-7508.

[41] Ye M, Wang Z, Lan X Y, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]. Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 1092-1099.

[42] Feng Z X, Lai J H, Xie X H. Learning modality-specific representations for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2020, 29: 579-590.

[43] Zhu Y X, Yang Z, Wang L, et al. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386: 97-109.

[44] Ye M, Shen J B, Crandall D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]. Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 2020: 229-247.

[45] Hao Y, Wang N, Gao X, et al. Dual-alignment feature embedding for cross-modality person re-identification[C]. Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019: 57-65.

[46] Wei Z, Yang X, Wang N, Gao X, et al. Flexible body partition-based adversarial learning for visible infrared person re-identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4676-4687.

[47] Zhao Y B, Lin J W, Xuan Q, et al. HPILN: a feature learning framework for cross-modality person re-identification[J]. IET Image Processing, 2019, 13(14): 2897-2904.

[48] Ye M, Lan X, Wang Z, et al. Bi-Directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 407-419.

[49] Liu H, Cheng J, Wang W, et al. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification[J]. Neurocomputing, 2020, 398: 11-19.

[50] Liu J, Sun Y, Zhu F, et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 19366-19375.

[51] Wang G A, Zhang T Z, Cheng J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea,

2019: 3622-3631.

[52] Kniaz V V, Knyaz V A, Hladuvka J, et al. ThermalGAN: multimodal color-to-thermal image translation for person re-identification in multispectral dataset[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 606-624.

[53] Li D, Wei X, Hong X, et al. Infrared-visible cross-modal person re-identification with an x modality[C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 4610-4617.

[54] Zhang Y K, Yan Y, Lu Y, et al. Towards a unified middle modality learning for visible-infrared person re-identification[C]. Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 788-796.

[55] Liu H J, Ma S, Xia D X, et al. SFANet: a spectrum-aware feature augmentation network for visible-infrared person reidentification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 1958-1971.

[56] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[57] He J C, Li L, Xu J C. Approximation properties of deep ReLU CNNs[J]. Research in the Mathematical Sciences, 2022, 9(3): 38-39.

[58] Nguyen D T, Hong H G, Kim K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

[59] 石跃祥, 周玥. 基于阶梯型特征空间分割与局部注意力机制的行人重识别[J]. 电子与信息学报, 2022, 44(01): 195-202.

[60] Wei Z Y, Yang X, Wang N N, et al. Dual-adversarial representation disentanglement for visible infrared person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2186-2200.

[61] Ye M, Shen J, Lin G, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[62] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]. Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 7794-7803.

[63] Radenovic F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655-1668.

[64] Tian X, Zhang Z, Lin S, et al. Farewell to mutual information: variational distillation for cross-modal person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021: 1522-1531.

[65] Liu H, Tan X, Zhou X. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J]. IEEE Transactions on Multimedia, 2021, 23: 4414-4425.

[66] Wang Z, Wang Z, Zheng Y, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019: 618-626.

[67] Wu A C, Zheng W S, Gong S G, et al. RGB-IR person re-identification by cross-modality similarity preservation[J]. International Journal of Computer Vision, 2020, 128(6): 1765-1785.

[68] Xu X H, Liu S, Zhang N, et al. Channel exchange and adversarial learning guided cross-modal person re-identification[J]. Knowledge-Based Systems, 2022, 257(4): 109883.

[69] Xiang X Z, Lv N, Yu Z T, et al. Cross-modality person re-identification based on dual-path multi-branch network[J]. IEEE Sensors Journal, 2019, 19(23): 11706-11713.

[70] Choi S, Lee S, Kim Y, et al. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020: 10254-10263.

[71] Basaran E, Gokmen M, Kamasak M E. An efficient framework for visible-infrared

cross modality person re-identification[J]. Signal Processing: Image Communication, 2020, 87: 115933.

[72] Zhang S Z, Yang Y F, Wang P, et al. Attend to the difference: cross-modality person re-identification via contrastive correlation[J]. IEEE Transactions on Image Processing, 2021, 30: 8861-8872.

[73] Fan X, Jiang W, Luo H, et al. Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal person re-identification[J]. Visual Computer, 2022, 38(1): 279-294.

[74] Zhang Y K, Wang H Z. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023: 2153-2162.

[75] Lu Y, Wu Y, Liu B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 13376-13386.

[76] Hao X, Zhao S Y, Ye M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 16383-16392.

[77] Lu H, Zou X Z, Zhang P P. Learning progressive modality-shared transformers for effective visible-infrared person re-identification[C]. Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 2023: 1835-1843.

[78] Liu J N, Wang J L, Huang N C, et al. Revisiting modality-specific feature compensation for visible-infrared person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 7226-7240.

[79] Wu Q, Dai P Y, Chen J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 2021: 4330-4339.

[80] Zeng X L, Long J. Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification[J]. Expert Systems with

Applications, 2023, 215: 119307.

[81] Zhong Z, Zheng L, Cao D L, et al. Re-ranking person re-identification with k-reciprocal encoding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 3652-3661.

致谢

行文至此落笔为终，二十余载求学路，或顺利或坎坷。回首这漫长而短暂的研究生旅程，曾笑过、哭过、憧憬过也迷茫过，目光所及，皆是回忆。不知不觉已走过三年，在安工程的学习生活使我受益匪浅。纵有万般不舍，仍心存感激。借此，向我的家人、老师、同学和朋友们致以最真诚的感谢。

桃李不言，下自成蹊。首先我要感谢我的导师王凤随教授，研究生生涯很幸运成为王老师的学生。老师求真务实的科研作风、严以律己的人生态度以及温和谦虚的处事风格深深影响着我，使我受益颇多。言传不如身教，感谢您三年以来对我的栽培，教我专业能力、受以学术素养。从论文的选题、构思、撰写、修改和定稿，离不开老师悉心的指导，一遍遍地审阅、以及逐字逐句地斟酌。每每遇到困难、迷茫无所时，与老师交流后总能扫清我心中忧虑，豁然开朗。遇到一位认真负责的好老师不易，再次感谢王老师三年的淳淳教诲与帮助。祝愿老师万事顺遂，喜乐平安。

父母之爱子，则为之计深远。感谢我的爸爸，妈妈和哥哥。感谢您们不求回报的付出，尊重我成长道路上的每一次选择，做我最坚实的后盾，让我有义无反顾向前地勇气。庆幸自己生活在幸福的一家四口，有一个照顾我许多的哥哥，虽幼时经常打闹拌嘴，但也陪伴我长大让我的童年从未感受孤独。感谢家人永远尊重我，支持我，疫情在家备考给我无微不至的照顾，让我心无旁骛地学习，给我幸福温暖的家庭环境，是我永远坚定的避风港。希望时间走的慢一些，让我陪伴你们的时间长一点，以后努力成为你们的依靠。祝愿我的家人永远身体健康，幸福平安。

岁月虽轻浅，时光亦潋滟。特别感谢 404 课题组的全体同学。感谢师兄师姐对我的照顾，在课题研究上提供细心细致的帮助与支持，在学习中给我建议，在撰写论文给我指导。感谢同门们互帮互助一起解决实验困难，共同奋斗，为我的实验生活增添了许多回忆。感谢实验室所有小伙伴三年的陪伴，使我的研究生学习生活过得非常充实快乐。愿未来万里鹏程，一切顺利！

浅喜似苍狗，深爱如长风。感谢我的男朋友束勤平，从大学到研究生，这一路走来你给予了我无数的包容与关怀。很幸运与你相遇，很庆幸这一路有你

的陪伴，互相见证彼此从青涩懵懂到成熟的蜕变。转眼间我们都即将毕业，回望备考期间我们一同学习，互相鼓励，收到同一所学校录取通知书的欣喜在心中依然清晰。感谢你一直的支持与鼓励，感谢在读研三年的起起落落中有你疏导我的坏情绪，做我的精神支柱。感谢你坚定不移的选择，给予我无限的爱与包容。人生路漫漫，有你陪伴的日子平凡且浪漫，愿未来我们携手共进退，一起奔赴美好的未来。

前程可奔赴，岁月可回首。感谢一路走来从未放弃，努力向前的自己。人生的路走的每一步都算数，二十余载求学路中每个刻苦拼搏的日夜，自我治愈、坚持的瞬间，都在不经息间化为我前行的力量，使我收获坚韧、成熟与责任。愿自己能一直保持热枕，山高路远，继续前行！

最后，感谢所有参加本论文评审和答辩的老师们，谢谢你们的辛苦付出！

攻读学位期间取得的学术成果情况

1、发表学术论文情况：

[1] 王路遥,王凤随,闫涛,陈元妹. 结合多尺度特征与混淆学习的跨模态行人重识别[J/OL]. 智能系统学报, 1-11. http://kns.cnki.net/kcms/detail/23.1538.TP.20240312.1048.002.html.

[2] 王路遥,王凤随,陈元妹. 多分支融合变分细化蒸馏的跨模态行人重识别[J/OL]. 重庆工商大学学报(自然科学版), 1-9. http://kns.cnki.net/kcms/detail/50.1155.N.20230427.1734.002.html.

[3] 陈元妹,王凤随,王路遥. 细化特征引导对抗性解纠缠学习的无监督行人重识别[J/OL]. 电子测量与仪器学报, 1-9. http://kns.cnki.net/kcms/detail/11.2488.TN.20240523.1630.004.html.

[4] 陈元妹,王凤随,钱亚萍,王路遥. 基于特征细化的多标签学习无监督行人重识别[J].浙江理工大学学报(自然科学), 2023, 49(06): 755-763.

2、参与科研项目情况：

（1）安徽省自然科学基金(2108085MF197)

（2）安徽高校省级自然科学研究重点项目(KJ2019A0162)

（3）安徽工程大学国家自然科学基金预研项目(Xjky2022040)

尚德教学 唯实惟新